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 y Tomás. Quiero agradecer también a mis compañeros de grupo y oficina, Lean, Yann, Javier, 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

Introduction

The first years of proton-proton collisions at a centre of mass energy of 7 TeV delivered by the Large Hadron Collider and recorded by the ATLAS experiment have provided data to explore quantum chromodynamics (QCD) at scales never reached before. Precision measurements of strong interactions are interesting in their own right, but, in addition, QCD provides one of the main backgrounds to many New Physics measurements; furthermore, it is also through tests of QCD that New Physics may be discovered.

Due to QCD confinement the experimental signature of quarks and gluons are not the quarks and gluons themselves but a spray of "colorless" hadrons, that we call *jets*. Hadronic jets are a fundamental ingredient for precision tests of QCD: understanding and measuring their performance is crucial in the LHC environment. A wide range of physics signatures, within the Standard Model (SM) and Beyond the Standard Model (BSM) predictions, contain jets originating from bottom (b) quarks. The ability to identify jets containing b-hadrons, the product of the hadronization of b-quarks, is therefore important for the high-p_T physics program of the ATLAS experiment.

b-tagging algorithms rely on the relatively long decay length of b-hadrons

that gives rise to large impact parameter tracks and displaced decay secondary vertices; or on the presence of a soft lepton within the jet, the product of the semileptonic b-decay. These algorithms, however, do not provide information on the number of b-hadrons within the jet. In particular, they tag "merged" jets containing a $b\bar{b}$ pair, with no net heavy flavour, which do not correspond to the intuitive picture of a b-jet as a jet containing a single b-quark or antiquark.

The ability to single out merged b-jets has several applications. The measurement of the QCD bottom production if of great importance due to the correspondence between parton level production and the observed hadron level, and its potential to provide information on the b-quark parton distribution function. The theoretical calculation of the inclusive b-jet spectrum presents rather large uncertainties ($\sim 50\%$), considerably larger than those for the light jet inclusive spectrum ($\sim 10 - 20\%$) [1]. It is found that the largest uncertainties are associated to the production channel known as "gluon splitting" (GSP), where a gluon from the hard scatter decays into a close-by $b\bar{b}$ pair, that a jet clustering algorithm often classifies within the same jet. This channel receives a strong enhancement from collinear logarithms spoiling the convergence of the perturbative series. An improvement in the accuracy of the theoretical predictions could be achieved by not including in the production cross-section the contribution from double b-hadron jets, which in QCD are produced $\sim 95\%$ of the time by the GSP channel.

Efficient tagging of merged b-jets can provide an important handle to understand, estimate and/or reject b-tagged backgrounds to SM analyses at the LHC that rely on the presence of single b-jets in the final state, such as top quark physics (either in the $t\bar{t}$ or the single top channels) or associated Higgs production ($WH \to \ell\nu b\bar{b}$ and $ZH \to \nu\nu b\bar{b}$). These processes suffer from

backgrounds that can be in part removed by a merged b-jet tagger. Jets containing a single b-quark or antiquark also enter in many BSM collider searches, notably because b-quarks are produced in the decays both of heavy SM particles (top quarks, the Z boson and the Higgs boson), and of particles appearing in proposed extensions of the SM. The ability to distinguish single b-jets from jets containing two b-hadrons is thus here of wide application to reduce SM backgrounds giving rise to close-by $b\bar{b}$ pairs.

There are two possible strategies to attempt to identify b-jets containing two b-hadrons in hadronic collisions. One of them, implemented at the CDF experiment at Fermilab [2], relies on the direct reconstruction of the two b-decay secondary vertices. This allows the measurement of the angular separation between the b-hadrons, but suffers from the low efficiency of a double b-tag requirement plus additional reconstruction inefficiencies at small angular separation between the two b-hadrons. In this thesis we develop for the first time an alternative method that does not rely on explicit vertex finding, but exploits the substructure differences between single and merged b-jets, combining them in a multivariate analysis. 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 in 2011.

The thesis is organized as follows: Chapter **??** describes the theoretical framework, with emphasis in the theory of the strong interactions and the aspects that are important for the understanding of the hadronic final state in hadronic collisions. The LHC and the ATLAS detector are described in Chapter **??**, together with a summary of the experimental conditions during the 2011 data taking. Chapter **??** details how jet reconstruction and calibration are performed at ATLAS and describes the procedure for the

identification of b-quark jets. Chapter **??** presents the analysis of jet shape and substructure variables for the discrimination between single and double b-hadron jets. The validation of the variables in 2011 data is also included. The construction of the multivariate discriminator and the discussion of the systematic uncertainties are presented in Chapter **??**. Chapter 2 details the technique used for the measurement of the fraction of double b-hadron jets in data and the associated systematic uncertainties. Finally, chapter 3 summarizes the results.

Chapter 2

Fraction of double b-hadron jets in QCD b-production

In this chapter we apply the newly developed $g- > b\bar{b}$ tagging tool to measure the fraction of merged b-jets in QCD b-jet production. The fractions are determined both for an inclusive b-jet sample with $|\eta| < 2.1$, and for exclusive samples enriched in single or in merged b-jet. The measured fractions are in excellent agreement within the experimental uncertainties with the theoretical prediction from a parton shower Monte Carlo simulation of hadronic collisions. The chapter is organized as follows. Section 2.1 introduces the concept of template fitting. Section 2.2 describes the statistical method used to perform the template fits. BLABLABLABLABLA

2.1 Introduction

The $g \rightarrow b\bar{b}$ tagger developed and described in the previous chapters produces for every b-tagged jet a number between 0 and 1, the double b-hadron likelihood (LL). The closer this number if to 1 (0), the more likely the b-

tagged jet is single (merged). When used as a tagger, a working point (Wp) is chosen so that if LL≥Wp the jet is flagged as single. The value of the Wp is chosen as a compromise between good efficiency (the lower the Wp, the higher the probability that an actual single b-jet will not be missed by the tagger), and rejection power (the higher the Wp the higher the probability that a non-single b-jet will not be incorrectly flagged as single). Depending on the necessities of the particular analysis, an appropriate Wp is to be chosen form the plot in Figure **??**, and in particular the performance results presented in Chapter **??** correspond to 50% and 60% efficiency working points, a reasonable choice.

However, the values of LL in a given sample offer more information than just a jet-by-jet tagger: the distribution of LL allows to measure the composition of the particular sample. In effect, a b-tagged jet has a certain probability to actually originate from the hadronization of a:

- b: b-quark
- $b\bar{b}$: gluon splitting into a $b\bar{b}$ pair
- c: c-quark
- $c\bar{c}$: gluon splitting into a $c\bar{c}$ pair
- ℓ: light parton (u, d, s quarks, or a g not splitting into heavy flavor pair).

The expected distribution of LL is different for each of the five cases. This is illustrated in Figure **??**, *which plots LL for each hypothesis for jets in two* p_T *ranges (*$40 < \mathrm{p}_T \leq 60$ *GeV and* $200 < \mathrm{p}_T \leq 270$ *GeV). These distributions are henceforth called "templates".* One can determine the composition of a given sample by measuring the values of LL of the b-tagged jets and

estimating the fractions needed from each of the templates to accurately describe the experimental LL distribution. This process is known as "template fitting".

The shape of the templates in Figure **??** *can be intuitively understood. The b and bb templates behave as expected, respectively peaking at high and low values of LL. The c and cc templates resemble their b counterparts. This was to be expected, given that a c-quark fragments mainly into into D-mesons which have a measurable* $c\tau$ *of* $\tilde{3}00\mu m$*, although shorter than the* $c\tau$ *of* $\tilde{5}00\mu m$ *corresponding to the B-mesons produced in b-quark fragmentation. Single and merged c-jets are thus the main background to b- and bb-jets. The template of light jets, on the contrary, do not contain large decay-length secondary vertices, and its distribution is driven by..*

2.2 Unbinned maximum likelihood fits

The analysis of experimental data often involves the estimation of the composition of a sample, based on Monte Carlo description 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 the parameters, $F(\vec{x}, \vec{p})$. We choose the PDF based on some hypothesis 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 data binned into a histogram, one approach is to use a least-squares fitting technique to estimate the parameters. They are adjusted to

minimize

$$\chi^2 = \sum_i \frac{(d_i - f_i)^2}{d_i} \tag{2.1}$$

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

$$f_i = N_D \sum_{j=1}^{m} p_j \cdot a_{ji}/N_j \tag{2.2}$$

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

This χ^2 assumes that the distribution for d_i is Gaussian and that a_{ij} has no uncertainty; it is of course Poisson, but the Gaussian $N(\mu = d_i, \sigma = \sqrt{d_i})$ 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 2.1 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

$$e^{-f_i} \frac{f_i^{d_i}}{d_i!} \tag{2.3}$$

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

$$\mathcal{L} = \prod_{i=1}^{n} e^{-f_i} \frac{f_i^{d_i}}{d_i!}. \tag{2.4}$$

This accounts correctly for the small numbers of data events in the bins. It is often referred to as a "binned maximum likelihood" fit. Actually 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 exists, see Ref. [3]. The effects of finite MC data size can be considered small for MC samples ten times larger than the data sample.

The technique of binned maximum likelihood fit is fast and analytical. Unfortunately this method only works satisfactory with weighted events if the weights do not differ very much [3]. We observed that the obtained uncertainties were unnaturally large. This was traced to the use of events with rather different weights. We had thus to move to a different, more general, technique, an "unbinned maximum likelihood fit", which allows arbitrary weights and has the further advantage of using all the information contained in the data sample; although it is not analytical but numerical and iterative.

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

$$\mathcal{L}(\vec{x};\vec{p}) = \prod_{k=1}^{N} F(x_k;\vec{p}) \tag{2.5}$$

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

$$\mathcal{L} = \prod_{k=1}^{N} \sum_{j=1}^{m} n_j \mathcal{P}_j(x_k) \tag{2.6}$$

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

The fits were performed in this thesis by means of the RooFit Toolkit for data modelling [4]. Performing a fit consists of minimizing the negative log-likelihood of a PDF calculated over the data set

$$-\log \mathcal{L}(\vec{p}) = \sum_{k} F(\vec{x}_k;\vec{p}) \tag{2.7}$$

with respect to the model's parameters. The RooFitTools package uses the MINUIT[5] 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,

$$M(x) = f_{sig} \cdot S(x) + (1 - f_{sig}) \cdot B(x), \tag{2.8}$$

or generically for N components:

$$S(x) = c_0 \cdot F_0(x) + c_1 \cdot F_1(x) + ... + c_{n-1}F_{n-1}(x) + (1 - \sum_{i=1}^{n-1} c_i)F_n(x) \tag{2.9}$$

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.

2.3 Fitting MC templates to data

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 fitted 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 hadrons 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 2.2). A separated fit

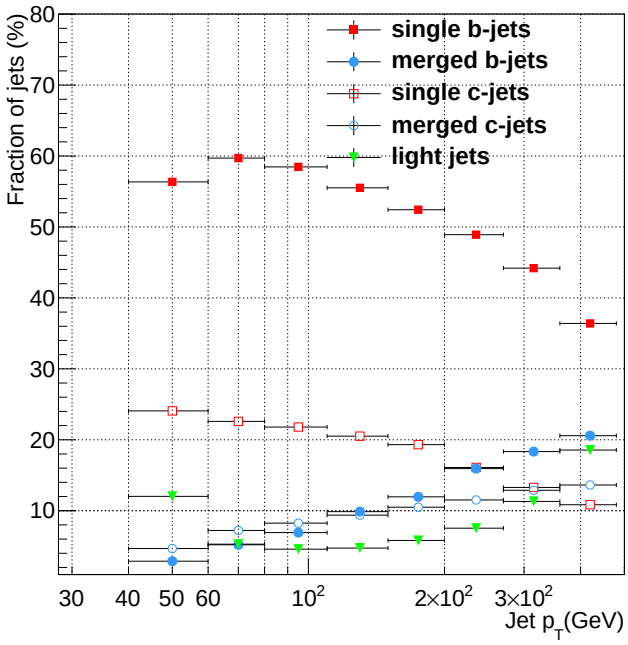

Figure 2.1: Theoretical predictions of the fractions of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets in b-tagged PYTHIA QCD sample.

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,

$$F(x) = n_s \cdot S(x) + n_m \cdot M(x) + n_l \cdot L(x) + n_{sc} \cdot S_c(x) + n_{mc} \cdot M_c(x) \quad (2.10)$$

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 fractions derived in* PYTHIA *as a function of the jet* p_T *are shown in Fig. 2.1.*

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 2.1. Examples of this set of fits are displayed in Figures 2.2 and 2.3.

Jet p_T	single b-jet			merged b-jet			light jet		
(GeV)	n_s	pred.	pull	n_m	pred.	pull	n_l	pred.	pull
40 - 60	62±3	56	0	3±1	3	0	4±4	12	0
60 - 80	62±1	60	0	5.2±0.4	5	0	2±2	5	0
80 - 110	57±1	58	0	8.5±0.4	7	0	3±2	5	0
110 - 150	55±2	56	0	13±1	10	0	1±4	5	0
150 - 200	53±3	52	0	15±1	12	0	0±4	6	0
200 - 270	53±5	49	0	17±1	16	0	-1±7	8	0
270 - 360	48±3	44	0	19±1	18	0	4±4	11	0
360 - 480	39±5	36	0	21±1	21	0	15±6	19	0

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

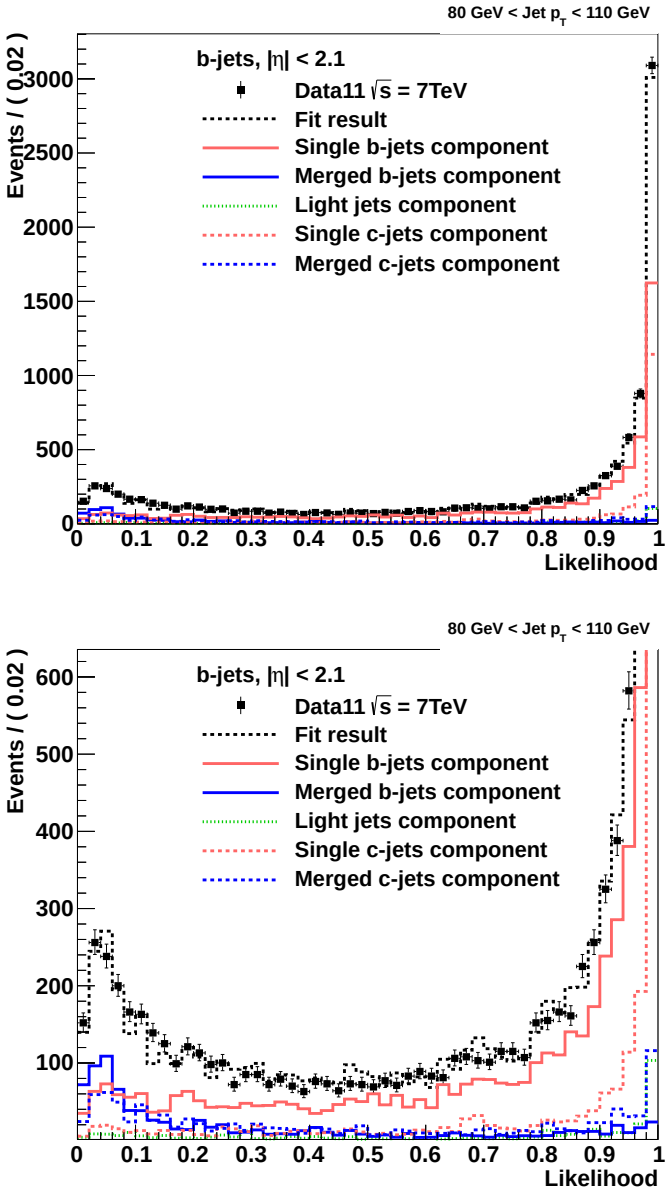

Figure 2.2: Example result of a template fit to the likelihood distribution in data. The fit is shown for jets with p_T between 80 GeV and 110 GeV, in full scale (top) and zooming the vertical scale, to better display the flavour content of the data (bottom). Uncertainties shown are statistical only.

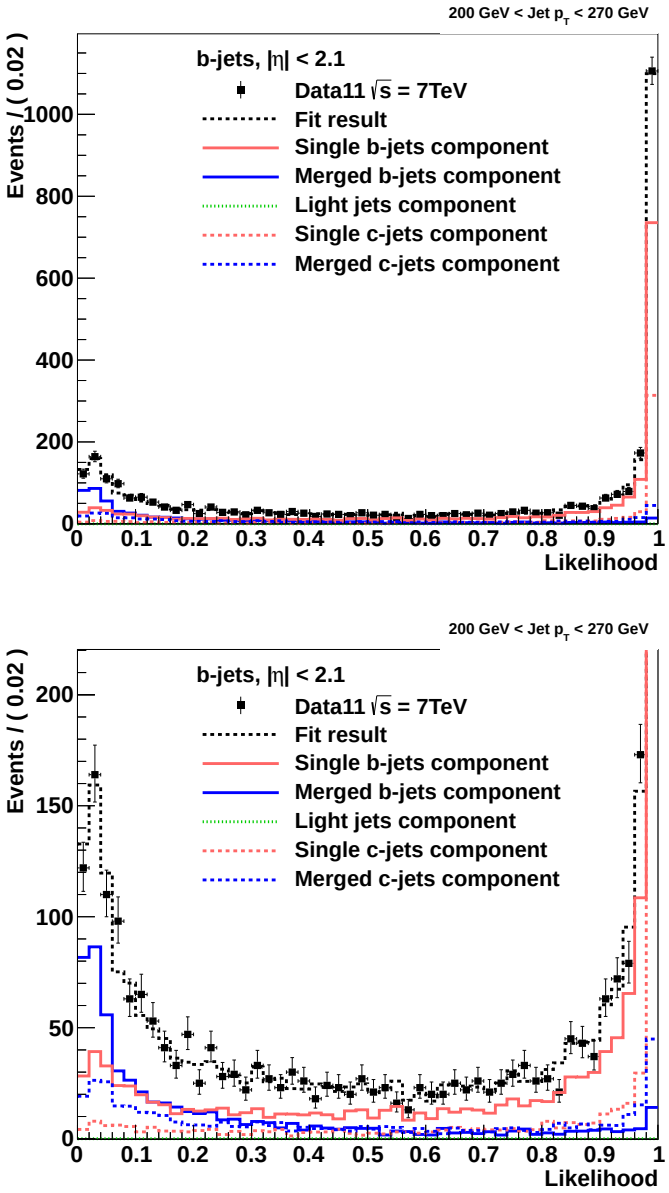

Figure 2.3: Example result of a template fit to the likelihood distribution in data. The fit is shown for jets with p_T between 200 GeV and 270 GeV, in full scale (top) and zooming the vertical scale, to better display the flavour content of the data (bottom). Uncertainties shown are statistical only.

2.4 Systematic uncertainties

The systematic uncertainties affecting the method are mainly those that change the shape of the likelihood templates 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 uncertainty in the track reconstruction efficiency the procedure described in Section **??** was followed. New likelihood templates 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 templates.

The impact of the uncertainty in the knowledge of the flavour relative c/b and $c\bar{c}/b\bar{b}$ fractions in the simulation was evaluated by separately changing these ratios by 20%. The variation in ratio $c\bar{c}/b\bar{b}$ 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 $b\bar{b}$- and $c\bar{c}$-flavoured components can be obtained, we are, in reality, measuring the fraction of merged $b\bar{b} + c\bar{c}$ together. The same result is observed if changing the single c/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%
heavy flavour fraction	negligible

Table 2.2: Contributions to the systematic uncertainties affecting the template fitting to experimental data.

2.5 Enriched samples in single and merged b-jets

The data sample employed in the analysis is ∼50% pure in single b-jets, according to the measurements described in Section 2.3. Having a purer data sample of single or merged b-jets would facilitate the validation of the Monte Carlo templates used for template fitting. To this end we consider the Monte Carlo dijet sample to determine the purity than can be achieved by simple kinematic and b-tagging cuts, leaving reasonable statistics.

In a MC parton shower generator such as PYTHIA generator, single b-jets

are produced mainly via the Flavour Creation (FCR) and Flavour Excitation (FEX) processes. In the former two heavy quarks are produced in the hard scatter. The FEX one can be depicted as an initial state gluon splitting into a $b\bar{b}$ pair, where one of the b-quarks subsequently enters the hard scatter, i.e., there is one b-quark in the final state. Merged b-jets are produced 95% of the time by Gluon Splitting (GSP). In this process no heavy quarks participate in the hard scatter, but they are produced via subsequent $g \to Q\bar{Q}$ branchings (see Section **??**).

In the FCR process two single b-quarks are produced, which lead to two back-to-back b-jets. Succesfully tagging these events is a way to build a sample enriched in single b-jets. On the other hand, events with only one b-jet can be produced either in the FEX process, with only one central b-quark in the final state, or via GSP, where $b\bar{b}$ pairs produced at small angles can be reconstructed as a single b-jet. These two scenarios are more difficult to disentangle.

Purified sample in single b-jets

To help purifying the sample in single b-jets we can then use the b-tagging information. Jets in data events with exactly two b-tags, selected with MV1 tagging algorithm at its 60% working point, and satisfying the event and jet selection described in Section **??** compose the purified data set. In order to increase the statistics no requirement on the p_T of the b-tagged jets was initially imposed for the event selection.

Once the enriched sample is obtained, new likelihood fits are performed, for all p_T bin, utilizing the same MC templates as for the nominal data sample in order to evaluate their performance. The fractions of single b, merged b and light in data recovered, together with their statistical errors and the

PYTHIA MC predictions for each p_T bin are displayed in tables 2.3 to **??**. Examples of these fits are shown in Figures 2.4 and Figures 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 of enrichment.

Purified sample in merged b-jets

Events with only one b-tagged jet were selected for the purification of the data sample in double b-hadron jets. In order to reinforce this selection, a tight anti-b-tagging on any non-tagged jet in the event was implemented. The anti-b-tagging was performed by imposing, simultaneously, strict cuts on the b-tagging weights of the three supported (calibrated) taggers available:

- MV1: $w < 0.07$
- JetFitter: $w < -2$
- IP3D+SV1: $w < -2$

These weight values correspond to a MV1 tagging efficiency of more than 85%, and an efficiency for b-tagging of more than 80% for the JetFitter and IP3D+SV1 algorithms.

The idea behind the anti-b-tagging requirement is to eliminate FCR events for which one of the produced b-quark jets failed to be tagged. However, although a certain level of purification was expected, very little enrichment was achieved with this kind of selection.

In order to understand this behavior an additional selection was applied: b-tagged jets were required to be back-to-back. *Most of the b-tagged jets are back-to-back. Three topologies are expected to be left after the selection described above: FCR back-to-back single b-jets not tagged, FEX single b-jets*

back-to-back with light jets, and GSP merged b-jets, also back-to-back with light. In 90% of the cases the second jet is a light jet; consecuently, b-jets arising from FEX process constitute the irreducible background for selecting merged b-jets. Very little enrichment can be attained.

Jet p_T	single b-jet		
(Ge)	n_s	pythia prediction	pull
40 - 60	99±11	84	0
60 - 80	82±5	87	0
80 - 110	84±5	88	0
110 - 150	86±8	85	0
150 - 200	89±9	83	0
200 - 270	95±5	80	0
270 - 360	67±11	81	0
360 - 480	73±16	73	0

Table 2.3: Measured proportions (in percentage) of single b-jets in experimental data from 2011 run, enriched in single b-jets.

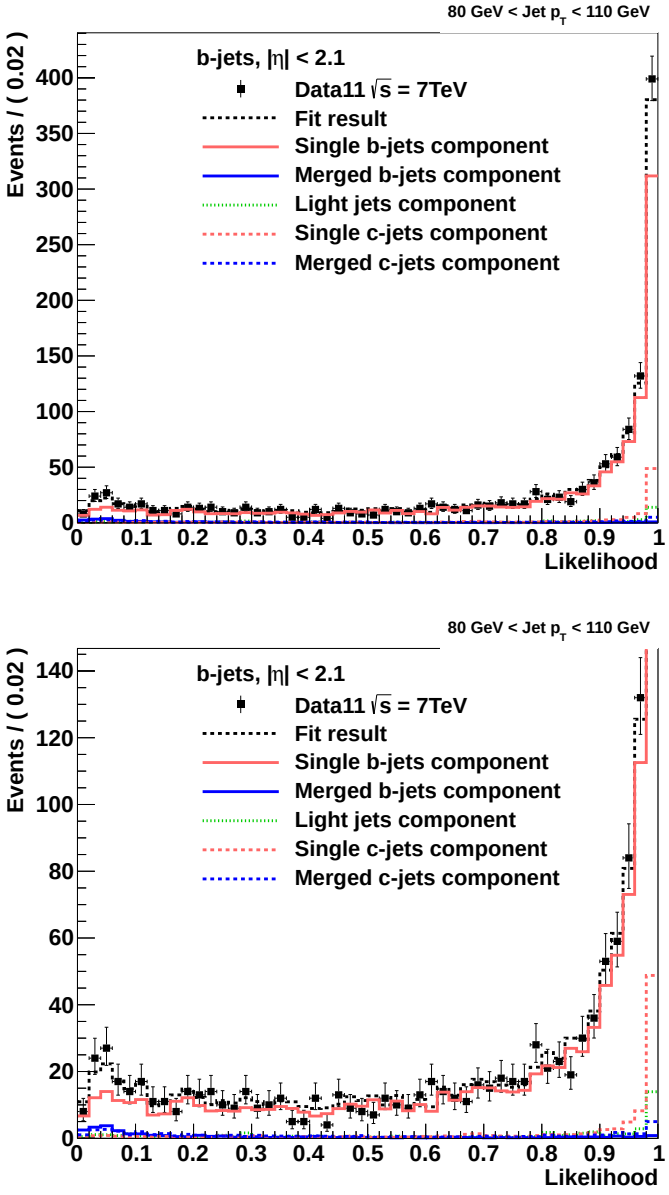

Figure 2.4: Example result of a template fit to the likelihood distribution in data enriched in single b-jets. The fit is shown for jets with p_T between 80 GeV and 110 GeV, in full scale (top) and zooming the vertical scale, to better display the flavour content of the data (bottom). Uncertainties shown are statistical only.

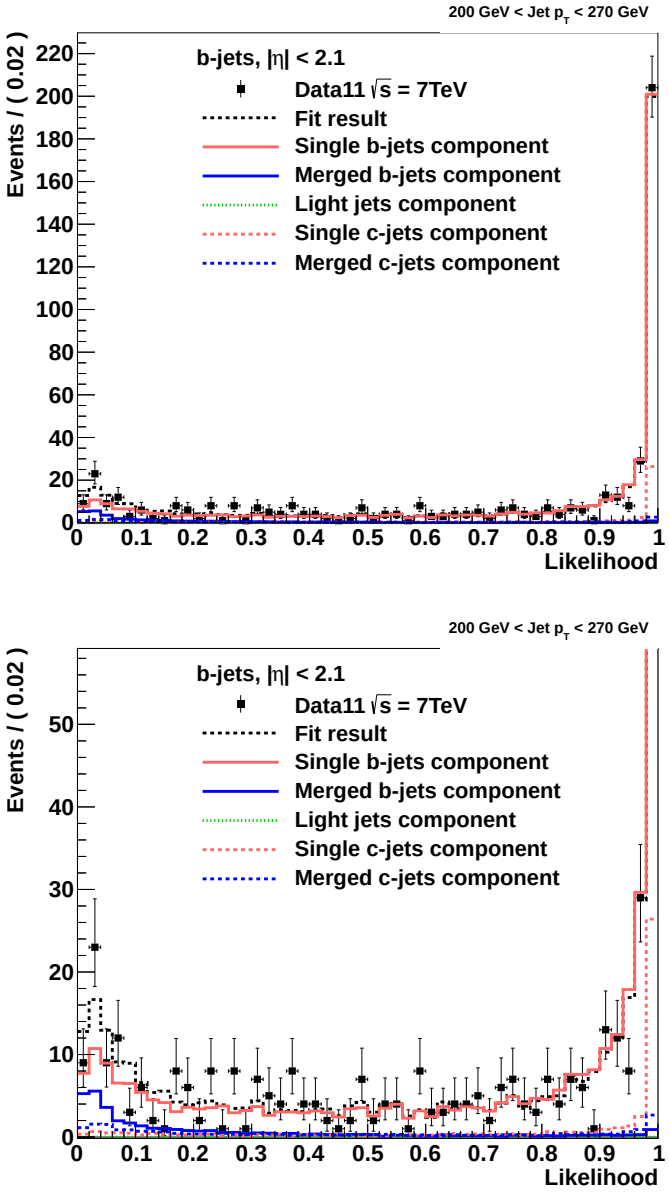

Figure 2.5: Example result of a template fit to the likelihood distribution in data enriched in single b-jets. The fit is shown for jets with p_T between 200 GeV and 270 GeV, in full scale (top) and zooming the vertical scale, to better display the flavour content of the data (bottom). Uncertainties shown are statistical only.

Jet p_T	merged b-jet		
(GeV)	n_m	pythia prediction	pull
40 - 60	3.3±0.5	3.1	0
60 - 80	6.1±0.5	5.9	0
80 - 110	9.2±0.5	8.2	0
110 - 150	14±1	12	0
150 - 200	16±1	15	0
200 - 270	19±1	19	0
270 - 360	20±1	22	0
360 - 480	23±1	23	0

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

Chapter 3

Summary and conclusions

In the course of the present thesis a new method allowing the identification of b-jets containing two b-hadrons was developed. The method exploits the expected kinematic differences between double b-hadron ("merged") jets and single b-jets, combining a set of discriminating variables in a multivariate classifier. The differences between single and merged jets originate in the two-subjet structure of merged jets which, in QCD, are expected to arise mainly from a gluon splitting into a close-by $b\bar{b}$-pair. Merged jets tend to have higher multiplicity and larger width. Several jet shape and substructure variables accounting for these envisaged characteristics were investigated in order to obtain the best single-merged discrimination. Due to the noisy environment of the hadron collisions at the LHC track-based variables were preferred over calorimeter variables.

The multivariate classifier was trained using simulated QCD events. A likelihood estimator was chosen among other multivariate methods for its robustness and simplicity. Based on discrimination power, correlation and pile-up dependence three input variables were selected for the tagger training: the jet track multiplicity, the track-jet width and the ΔR between the axes

of two k_t subjets in the jet. The peformance of the tagger in Monte Carlo events was studied in bins of the calorimeter jet p_T, achieving a rejection of merged jets of over 95% (90%) for a 50% single b-jet efficiency for jets with $p_T > 150$ GeV ($p_T > 60$ GeV). Several sources of systematic uncertainties in the merged b-jet rejection were evaluated for the 50% and 60% single b-jet efficiency working points, the most relevant being the tracking efficiency and the jet energy scale and resolution with an average performance variation of 4%, 5% and 5%, respectively. Other contributions such as pile-up or the uncertainties in the track momentum resolution and the b-jet tagging efficiency proved to be negligible.

The Monte Carlo distributions of the explored variables as well as the likelihood output were validated using experimental data corresponding to an integrated luminosity of 4.7 fb^{-1} recorded by the ATLAS experiment during 2011. The agreement between data and simulation is excellent.

Monte Carlo templates were used to fit the likelihood distribution in data in order to obtain the fraction of merged b-jets in the data sample. This measurement was performed by means of unbinned maximum likelihood fits. The systematic uncertainties that most affected the method were the uncertainties in the jet energy scale and the jet energy resolution, with an average variation of 2% each.

This tool provides a handle to investigate QCD $b\bar{b}$ production and to reduce backgrounds in Standard Model physics analyses that rely on the presence of single b-jets in the final state, such as top quark physics (either in the $t\bar{t}$ or the single top channels) or associated Higgs production ($WH \to \ell\nu b\bar{b}$ and $ZH \to \nu\nu b\bar{b}$). Jets containing a single b-quark or antiquark also enter in many BSM collider searches, the ability to distinguish single b-jets from jets containing two b-hadrons is thus here of wide application to reduce SM

backgrounds giving rise to close-by $b\bar{b}$ pairs.

In order to expand up the results presented here, and to make further advancements in the implementation of the tagger in physics analyses the following improvements should be made: the extension to non-isolated jets using the concept of ghost-particle matching and active area of a jet for track-to-jet association and labeling and the calibration of the tagger with data. Nontheless, the study presented in this thesis demonstrates that jet substructure variables can provide a good handle for gluon splitting identification in physics searches within ATLAS.

Bibliography

[1] S. Frixione and M.L. Mangano. Heavy quark jets in hadronic collisions. *Nucl.Phys.*, B483:321–338, 1997.

[2] CDF Collaboration. Measurements of Bottom Anti-Bottom Azimuthal Production Correlations in Proton-Antiproton Collisions at $\sqrt{s} = 1.8$ TeV. *Phys.Rev.D*, 71:38, 2005.

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

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

[5] James, F and Roos, M. Function minimization and error analysis. *CERN Computer Center Program Library D*, 506, 1983.