DELPHI Collaboration DELPHI 97-134 PHYS 721
August 11, 1997

Description and performances of the DELPHI multivariate flavour tagging

Ch. De la Vaissière[1], F. Martínez-Vidal[2]

Abstract

This note updates the description of the DELPHI multivariate hemisphere flavour tagging algorithm. It reviews the performances achieved for all the LEP 1 data, in particular after the upgrade in early 1994 of the microvertex detector including z readout. Tagging variables have been redefined incorporating two and three dimensional impact parameter reconstruction, transverse momenta, secondary vertices and invariant mass reconstruction. The variables used in the 1991-1993 periods have been revised according to this new definition in order to improve performances and to unify the tagging for the different data taking periods. Significant improvements have been obtained. The tagging in the two hemispheres has been found almost independent owing to a separated reconstruction of the primary vertex for each hemisphere.

[1]LPNHE, Laboratoire de Physique Nucleaire des Hautes Energies, Universités Paris VI et VII, Paris.
[2]IFIC, Centro Mixto Univ. of Valencia – CSIC, Valencia.

Contents

1 Introduction **1**

2 Tagging algorithm **1**

3 Track and event selection for tagging **2**

4 The data and Monte Carlo samples **4**

5 The hemisphere primary vertex finder **4**

6 Impact parameter reconstruction **7**
6.1 Signed impact parameter 8
6.2 Track-jet distance in space 9
6.3 The track helix linearization 11
6.4 Signed impact parameter in two dimensions 11
6.5 Impact parameter errors 12
6.5.1 2D impact parameter errors with respect to the hemisphere vertex . 12
6.5.2 3D impact parameter errors with respect to the hemisphere vertex 13
6.6 Impact parameter significance 14

7 Tracking tuning **15**

8 Probability of primary vertex decay products **15**

9 Search for secondary vertices **16**

10 Weights of B hadron decay products **18**

11 Definition of the multivariate tagging variables **20**
11.1 Variables from secondary vertex search 21
11.1.1 Secondary vertex counter (*SumNSV*) 21
11.1.2 Secondary vertex proper decay length (*SumDSV*) 21
11.1.3 Secondary vertex mass (*MaxMSV*) 21
11.2 Variables using B decay weights 21
11.2.1 Weighted mass (ω_{mass}) 22
11.2.2 Total weight 1 (Ω_1) 22
11.2.3 Total weighted $p_\perp$ ($\Omega_{p_\perp}$) 23
11.2.4 Total weighted p (Ω_p) 23
11.2.5 Total weight 2 (Ω_2) 23
11.3 Miscellaneous variables 23
11.3.1 Boosted sphericity ($\ln S$) 23
11.3.2 Normalized decay path (λ) 24
11.3.3 Sum of normalized track-jet abscissa (ξ) 26
11.3.4 Excluded particles (N_{exclu}) 26
11.3.5 Hemisphere primary vertex probability ($\mathcal{P}_H^+$) 26

12 Flavour confidences **26**

13 Combined multivariate/confidences tagging **27**

14 Tagging performances **30**

15 Hemisphere-hemisphere tagging correlation **31**

16 Giving access to the multitag **32**

17 Conclusions **34**

1 Introduction

The measurement of the partial hadronic branching ratio to $b\bar{b}$ pair quarks, R_b, requires an efficient and pure b quark tag associated with a good hemisphere independence in order to reduce and control as much as possible systematic errors. This has to be achieved with the complex mixture of $Z \to hadrons$ final states at LEP 1, so the experimental task is not too easy. The multivariate tagging in DELPHI [1], designed to satisfy these two requirements, was used for the R_b measurement with the 1991 to 1993 data [2, 3]. After the upgrade of the microvertex detector in spring 1994 a revision of the algorithm was needed. This is the subject of this note.

The note intends also to be a reference document to the R_b analysis, where the details on the variables used in multivariate technique can be found. For this reason it is long. An index would allow a reader interested to a specific topic to go to the relevant sections.

2 Tagging algorithm

The tagging algorithm is based on the large mass and relatively long lifetime of the b quark (typically 1.6 ps) and some event shape properties of their decays. All the available information is combined using multivariate techniques. The lifetime information exploits the large impact parameters of tracks coming from B decays together a search for secondary and tertiary vertices with their invariant masses. Finally, the lifetime information is combined with event shape properties of the B decays like large transverse momentum of the tracks with respect the jet axis and the boosted sphericity.

Tagging b quarks is based on reconstructing as precisely as possible the position of the primary Z boson decay, the track parameters of the outgoing particles with respect to the reconstructed vertex and applying an algorithm optimizing the use of this information.

For each single variable $\mathcal{Z}$, the probability $p_l(z_m)$ to observe a value z in the interval z_m for a hemisphere of flavour l is given by the content $y_l(z_m)$ of the corresponding bin m in the density distribution of this variable for flavour l:

$$p_l(z_m) = \frac{y_l(z_m)}{N_l^{tot}} \tag{1}$$

where N_l^{tot} is the total number of events in the l flavour distribution. The density distribution $y_l(z)$ is modelized by a *training sample* of simulated events that is different and tuned for each data set period [1]. Given the observed set $\left\{z^1_{m_1}, z^2_{m_2}, ..., z^N_{m_N}\right\}$, where the superscript corresponds to each variable and the subscript to particular bins, the probability that it comes from a given quark flavour uds, c and b is

$$\mathcal{P}_{uds} = \frac{3\prod_{\lambda=1}^{N} p_{uds}(z^\lambda_{m_\lambda})}{3\prod_{\lambda=1}^{N} p_{uds}(z^\lambda_{m_\lambda}) + \prod_{\lambda=1}^{N} p_c(z^\lambda_{m_\lambda}) + \prod_{\lambda=1}^{N} p_b(z^\lambda_{m_\lambda})} \tag{2}$$

$$\mathcal{P}_c = \frac{\prod_{\lambda=1}^{N} p_c(z^\lambda_{m_\lambda})}{3\prod_{\lambda=1}^{N} p_{uds}(z^\lambda_{m_\lambda}) + \prod_{\lambda=1}^{N} p_c(z^\lambda_{m_\lambda}) + \prod_{\lambda=1}^{N} p_b(z^\lambda_{m_\lambda})} \tag{3}$$

[1] In addition, to reduce statistical fluctuations Gaussian and exponential fits are performed to some tail distributions.

$$\mathcal{P}_b = \frac{\prod_{\lambda=1}^{N} p_b(z^{\lambda}_{m_\lambda})}{3\prod_{\lambda=1}^{N} p_{uds}(z^{\lambda}_{m_\lambda}) + \prod_{\lambda=1}^{N} p_c(z^{\lambda}_{m_\lambda}) + \prod_{\lambda=1}^{N} p_b(z^{\lambda}_{m_\lambda})} \tag{4}$$

respectively, being N the total number of variables used. The empirical factor 3 assigned to uds reflects the fact that this flavour is the sum of the three lighter flavours u, d and s which are taken together because their distributions are very similar. With this formulation the 5 flavours have the same weight.

This method of combining the probabilities may not be optimal. The individual probabilities used independently, are in fact correlated. Thus there is no statistically correct way to combine them, and several techniques could be tried. However, this choice was proven to be the best of several tried.

What counts in comparing flavours are ratios of probabilities or differences of their logarithms. For this reason we have introduced three estimators

$$\mathcal{L}_{uds} = \frac{2\ln\mathcal{P}_{uds} - \ln\mathcal{P}_c - \ln\mathcal{P}_b}{3} \tag{5}$$

$$\mathcal{L}_c = \frac{2\ln\mathcal{P}_c - \ln\mathcal{P}_{uds} - \ln\mathcal{P}_b}{3} \tag{6}$$

$$\mathcal{L}_b = \frac{2\ln\mathcal{P}_b - \ln\mathcal{P}_{uds} - \ln\mathcal{P}_c}{3} \tag{7}$$

called *flavour likelihoods*, which are the basis of the classification. The event can be classified according to the largest flavour likelihood (which is positive). The magnitude of the *best likelihood* is a sensitive indicator of the tag clarity. Based on this value, each tag can be subdivided into categories according to a set of given cuts.

3 Track and event selection for tagging

The starting point for flavour tagging is the selection of good hadronic Z decays according to the following criteria. In order to perform this selection, we have adopted the standard *TEAM 4* cuts of the DELPHI experiment. Firstly, charged particles are accepted if:

- their polar angle is between 20^o and 160^o;
- their track length is > 30.0 cm in the TPC;
- their momentum is > 200 MeV/c with a relative error less than 100%;
- their impact parameter (see section 6) relative to the interaction point is < 2.5 cm in the plane perpendicular to the beam direction and < 10.0 cm along the beam direction.

Events were selected by requiring:

- at least 5 reconstructed charged particles;
- the summed energy of the charged particles had to be larger than 12% of the center of mass energy;

- thrust axis satisfying $|\cos\theta_{thrust}| < 0.95$, where θ_{thrust} is the polar angle of the event thrust axis (section 5).

With these cuts the efficiency to select hadronic events was about 95% with all backgrounds (mainly from $\tau^+\tau^-$ pairs but also from $\gamma\gamma$ collisions) below 0.1%, without any significant bias in the flavour composition of the sample. Additional requirements on detector availability (provided by the slow control system) were required. No selection on the center of mass energy has been made.

The tagging is defined only from a subsample of physical *tight 2D* and *tight 3D* tracks required to have been produced near the interaction point. In addition to the *TEAM 4* cuts, *tight 2D* tracks have to satisfy the following conditions:

- hits in at least 2 of the 3 $R\phi$ layers of the VD;
- the $R\phi$ impact parameter (section 6) with respect to the main event vertex (section 5) less than 0.30 cm;
- the track was not associated to a reconstructed K^0, Λ or e^+e^- pair from photon conversion (see below).

Tight 3D tracks require further the following conditions:

- hits in at least 1 of the 2 z layers of the VD;
- the impact parameter with respect to the main event vertex in z less than 1.0 cm;
- no error code in the 3D impact parameter routine (section 6);
- the track-jet abscissa (section 6.2) less than 2.0 cm.

It happens that for a small fraction of the accepted events (around 0.1%) no tight tracks are found in none hemisphere. The event is then rejected because no tagging information is available in that case.

Finally, due to the limited angular acceptance of the microvertex detector an additional event polar angle acceptance cut is needed. A cut at 0.65 on $|\cos\theta_{thrust}|$ was imposed. The physical reason of this hard cut instead a softer cut (for instance at 0.75) is to reduce and control as much as possible hemisphere tagging correlations from VD edge effects. No additional cut on the number of jets in the event is performed. With all these cuts the global efficiency to select the hadronic events was about 60%.

As said above, selected tracks are required not to be associated to a reconstructed K^0, Λ or e^+e^- pair from photon conversion (V^0's). Candidate V^0 decays in hadronic events are found by considering all pairs of oppositely charged particles and then reconstructing the vertex. V^0 candidates are determined according to the standard DELPHI algorithm described in the first reference of [4]. There is no protection against short range Σ^+ and Σ^-. Also there is a small, but non zero probability that charged pions and kaons decay inside the beam pipe.

4 The data and Monte Carlo samples

The total number of accepted hadronic Z decays from the 1991 to 1995 runs of the LEP collider [2], before and after the angular acceptance cut, is summarized in table 1.

Table 1: Number of hadronic Z decays accepted for the analysis in each year of operation.

Year	1991	1992	1993	1994	1995	Total
Before $\mid \cos\theta_{thrust} \mid < 0.65$ cut	247277	691658	698557	1370354	664676	3672522
After $\mid \cos\theta_{thrust} \mid < 0.65$ cut	150635	421741	425796	828168	400482	2226822

Samples about three times the data statistics of $Z \to q\bar{q}$ events has been simulated using the Lund parton shower Monte Carlo JETSET 7.3 [5] and the DELPHI detector simulation DELSIM [6]. In addition, dedicated samples of $Z \to b\bar{b}$ events have been generated, with an equivalent statistics of also three times the data. The simulated events have been passed through the same analysis chain as the real ones. The parameters used in JETSET were optimized by DELPHI [7], in particular some parameters to which bottom and charm physics are sensitive:

- fragmentation function for heavy flavours, taken as Peterson et al. [8];
- the production fractions of weakly decaying charmed hadrons in $c\bar{c}$ events;
- the lifetimes of the charmed hadrons;
- the average charged decay multiplicities of the charmed hadrons;
- the production rates of b and c quarks via gluon splitting;
- the production rates of K^0's and hyperons.

5 The hemisphere primary vertex finder

A primary vertex fit serves to estimate the position of the interaction point. In a first step we determine an *event vertex* . Its purpose is to decide if a track originates from the production point region can be selected as a tight track as described in section 3. The position of the event vertex is computed using an iterative procedure which starts with all the charged particles of the event having an impact parameter in $R\phi$ with respect to the beam spot position less than 4.0 cm (very soft cut), by minimizing the full 3D least squares ansatz [9]:

$$\mathcal{M} = \sum_{j=1} \vec{\delta}_{a,j}^{\,T} \tilde{G}_j \vec{\delta}_{a,j} + \sum_{j=1} \left\{ \frac{(b_{x,j} - V_x)^2}{\sigma^2_{b_{x,j}}} + \frac{(b_{y,j} - V_y)^2}{\sigma^2_{b_{y,j}}} \right\} \tag{8}$$

In equation (8), $\vec{\delta}_{a,j}$ is the vector of closest approach in space of the track to the candidate vertex $\vec{V}$ and $\tilde{G}_j$ is the weight matrix of track j. Second term of (8) corresponds to the

[2]The used processings of the data are the last available at the moment when this work was written: 91F1, 92D2, 93C1, 94C2 and 95D2.

inclusion of the beam spot position ($b_{x,j}$,$b_{y,j}$) and dimensions ($\sigma^2_{b_{x,j}}$,$\sigma^2_{b_{y,j}}$) as a constraint of the vertex fit. This constraint is meaningful only in the $R\phi$ plane. At each iteration a search for the track with maximum contribution to the full 3D least squares ansatz above a threshold of 10.0 is done. If found, the track is removed and a new vertex fit is attempted until no track is removed. If no tracks are finally left, the beam spot position is used as estimate of the vertex. Since the beam spot position is used as a starting reference point, in principle all the tracks can be rejected from the fit. For these events the beam spot center is taken as the event main vertex and the covariance matrix corresponds to the beam spot size. The fraction of such events is around 1%.

The beam spot is defined as the interaction region of the electron and positron beams. To follow variations during the LEP fill, its position is determined for every cartridge written by the DAS corresponding to about 200 hadronic events. The x and y positions are found with typical uncertainties of about 9 μm and 4 μm respectively. The width along the x coordinate varies with time but a typical value is 100 to 120 μm with an error of 7 μm. The beam spot is small, which improves the accuracy of the event by event primary vertex fit and therefore the efficiency for tagging b quark events.

However, the fact that this primary vertex shares tracks from both hemispheres introduces correlations between the hemispheres:

- the lifetime of the B hadron on each side. If one B hadron has a long decay length, it will be almost certainly tagged. However, it will degrade the resolution of the primary vertex, making it less likely that the second B hadron will tag;
- the primary vertex error. If two hemispheres share a common primary vertex and if its error happens to be large, the B hadrons will be less likely to be tagged b;
- if the primary vertex is pulled towards one of the B hadrons (because it includes decay tracks), the decay range of that B hadron will be underestimated, while that of the other B will be overestimated.

These problems can almost be eliminated if a primary vertex is computed separately for each hemisphere. It should be remarked that the price to pay for this independence is a small decrease of tagging efficiency. However the reduction of hemisphere correlations has proven to be one of the main points for the R_b analysis [2, 3].

Back to back hemispheres are defined by classifying particles into two subsets using the event thrust axis. The thrust axis $\vec{T}$ is defined to maximize the sum [5]

$$\frac{\sum_a \mid \vec{p}_a \cdot \vec{T} \mid}{\sum_a \mid \vec{p}_a \mid} \tag{9}$$

where $\mid \vec{T} \mid = 1$. Index a runs over all the final state particles (neutral and charged) and $\vec{p}_a$ is the momenta of each particle. The maximal value found is known as event thrust. Particles are distributed into jets using the JADE algorithm [5] with $y_{cut} = 0.01$ and the jet direction was given by the jet thrust axis. Then particles are assigned to the hemisphere of the jet they belong to.

From this list of particles, an *hemisphere primary vertex* is now evaluated. Tracks with wrong associations to hits in the VD, from secondary decays of long lived particles or from interactions in the detector material, may spoil the reconstruction of the vertex.

To minimize the presence of these tracks, in a first step all the previously identified tight tracks of the hemisphere are used for the hemisphere vertex fit, taking as approximative solution the global event vertex previously computed. Then a selection of tracks is performed by requiring an $R\phi$ impact parameter less than 0.30 cm and less than 2.5 cm in z with respect to the vertex obtained in this first step. In the second step, with the selected tracks a new vertex fit is performed. If the fit probability of the full 3D least squares ansatz of equation (8) is less than 0.05, the particle with the most important contribution is removed, and a new vertex iteration is attempted. If no tracks are left in the fit (this happens on simulation less than of 4% of hemispheres), the event vertex is taken. From this fast algorithm the hemisphere vertex position, as well as the full covariance matrix, are determined.

Figure 1 shows the difference between the reconstructed and generated vertex positions in the x, y and z directions for light, charm and b hemispheres for the 1994 simulation. By comparison, table 2 summarizes the RMS of the obtained distributions for the 1994 and 1993 simulation. In 1994, for light quark events the RMS of the distribution in the x direction is about 60 μm and for b quarks it is around 125 μm; in the y direction it is around 10 μm for both, uds and b quarks. Therefore, the y primary vertex resolution is similar for uds and b quarks, because the tight beam spot constraint in that component. However compared to uds hemispheres, the x resolution for b quarks shows: a) higher RMS which is the result of the exclusion in the vertex fit of secondary tracks reducing the track multiplicity of the fit together a poorer beam spot determination (compared to the y component); b) larger tails because of the inclusion in that fit of some secondary tracks. In the z component similar arguments to the x component can be applied, with the additional consideration that the beam spot in z is not a real constraint in the vertex fit. Before 1994 the VD did not provide measurements of the z coordinate. Table 2 shows the factor about 8 of gain in z resolution for b hemispheres from 1993 to 1994, as consequence of the upgrade of the microvertex detector with z readout. In the x coordinate the resolution before 1994 is slightly poorer and it is similar for the y coordinate.

Table 2: RMS of the distributions of differences between the reconstructed and generated vertex positions in the x, y and z direction for light quarks, charm quarks and b quarks compared for 1994 and 1993 simulation.

Distribution	1994 Simulation	1993 Simulation
PVx-PVx(true) uds	56.6 μm	69.5 μm
PVx-PVx(true) c	73.8 μm	87.9 μm
PVx-PVx(true) b	125.3 μm	144.7 μm
PVy-PVy(true) uds	9.8 μm	9.9 μm
PVy-PVy(true) c	10.0 μm	10.0 μm
PVy-PVy(true) b	10.3 μm	10.3 μm
PVz-PVz(true) uds	85.2 μm	783.0 μm
PVz-PVz(true) c	99.0 μm	803.5 μm
PVz-PVz(true) b	147.4 μm	875.0 μm

Figure 2.a-c shows the differences between the reconstructed primary vertex and the beam spot. For the 1994 data, the RMS of the x, y and z distributions are 133.1 μm, 3.3 μm and 7050 μm respectively, compared with 130.9 μm, 3.0 μm and 7109 μm obtained

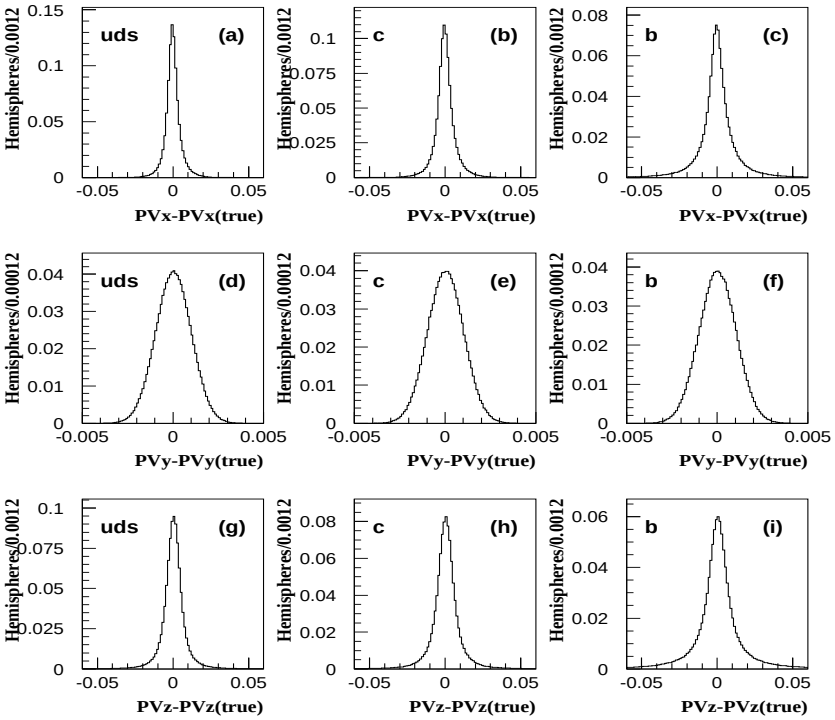


Figure 1: Difference between the reconstructed and generated hemisphere vertex positions in the x, y and z directions for light quarks, charm quarks and b quarks in the 1994 simulation. Horizontal scale is in cm.

from the Monte Carlo sample. Figure 2.d-f shows also the error obtained from the hemisphere vertex fit. The large tail of the z component is due main to bad measured tracks in z and the poor beam spot determination in that component.

Finally, figure 3 shows the differences between the positions of the two hemispheres in data and simulation for 1994, and table 3 summarizes the RMS of the distributions. As previously, the x and z distributions have larger tails because of the inclusion of secondary tracks and the poorer beam spot constraint.

Table 3: RMS of the distributions of differences between the two reconstructed hemisphere vertex positions in the x, y and z direction for 1994 simulation and data.

Distribution	1994 Simulation	1994 Data
PVx1-PVx2	91.1 μm	90.3 μm
PVy1-PVy2	3.8 μm	4.3 μm
PVz1-PVz2	155.4 μm	161.6 μm

6 Impact parameter reconstruction

Since the experimental track precision in the three spatial dimensions are comparable (when VD hits in $R\phi$ and z have been associated to the track), normal 3D metric for

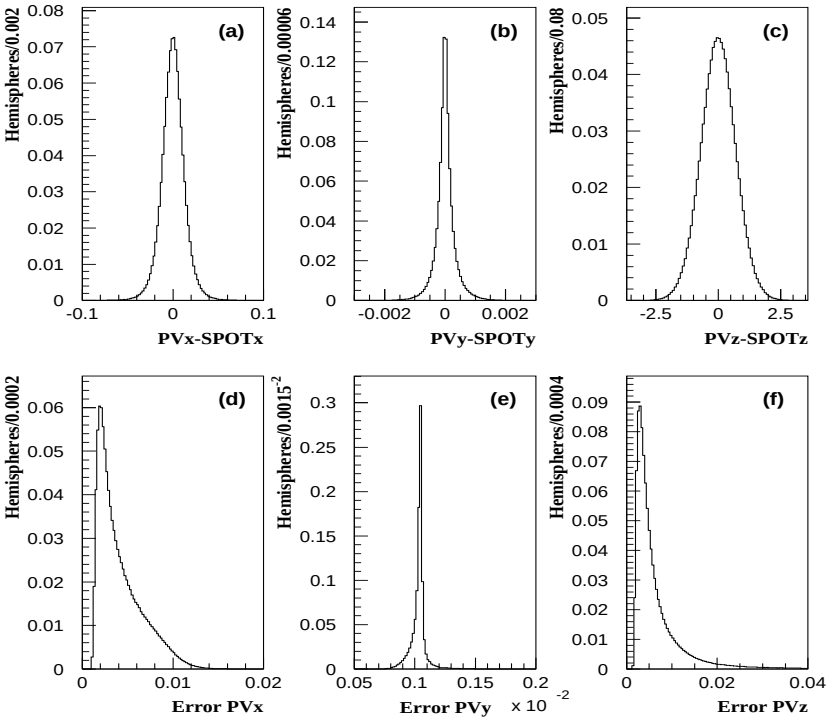


Figure 2: Vertex position with respect to the beam spot and errors for the 1994 data. Horizontal scale is in cm.

impact parameter reconstruction can be used. It has been found that weighting the $R\phi$ and z coordinates to take into account the differences in accuracy do not bring significant improvements with respect to the standard 3D calculations.

Conceptually, the impact parameter is the distance of closest approach between a track and the interaction point. The trajectory is represented by an helix. The usual convention is to take for the starting point of the helix a point $\vec{P}_0$ which is the perigee with respect to the origin of the DELPHI reference frame. The trajectory is defined through the usual 5 helix parameters $(h_0^{xy}, \Delta_z^0, \theta_0, \phi_0, 1/\rho)$ taken at perigee $\vec{P}_0$. The coordinates of $\vec{P}_0$ are therefore $(h_0^{xy} \sin\phi_0, -h_0^{xy} \cos\phi_0, \Delta_z^0)$. The point defines an origin on the helix. The position of another point of abscissa s (path length of the helix) can be calculated directly knowing the direction $\vec{T}_0$ (defined by ϕ_0 and θ_0) at $\vec{P}_0$ and the curvature $1/\rho$.

We approximate the interaction point by the hemisphere primary vertex, represented on figure 4 by the point $\vec{V}$. The value of $s = (\vec{V} - \vec{P}_0) \cdot \vec{T}_0$, defines a new point $\vec{P}_a$ (see section 6.3) which is the point of closest approach of the track with respect to the interaction point $\vec{V}$. The *3D impact parameter magnitude in space* will be therefore $\delta_a = \mid \vec{P}_a - \vec{V} \mid$.

6.1 Signed impact parameter

A b quark decay along its line of flight, downstream its production point. The purpose of attributing a sign to the impact parameter is to recognise that situation. One assumes that the direction $\vec{J}$ of the most energetic jet represents the quark direction. The line of

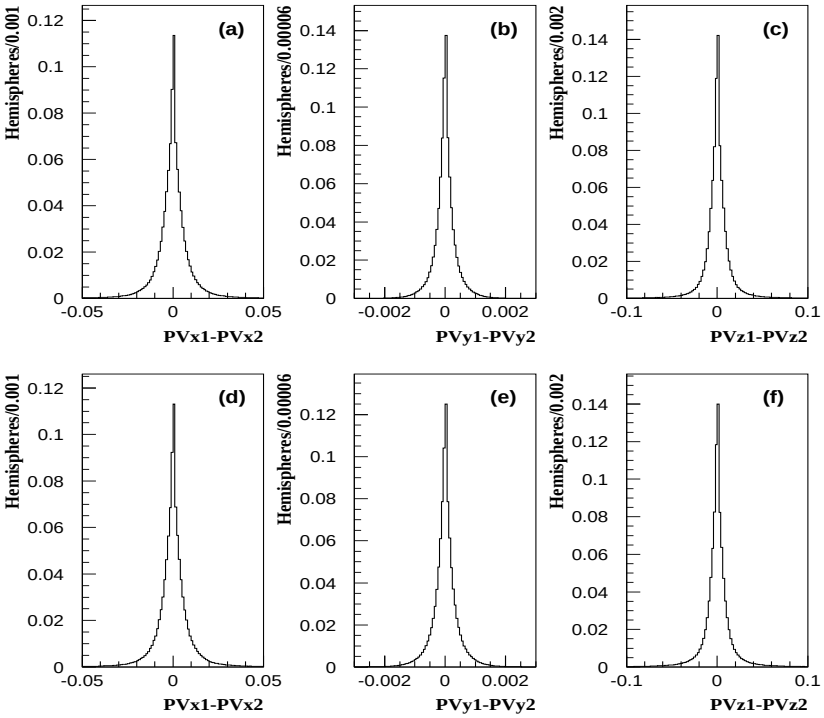


Figure 3: Difference between the two hemisphere vertex positions in simulation (a,b,c) and data (d,e,f) for 1994. Horizontal scale is in cm.

direction $\vec{J}$, attached to the vertex $\vec{V}$, approximates the line of flight of the quark. A first interesting quantity is the *projected impact parameter on the jet axis*

$$q_J = \overrightarrow{P_a V} \cdot \vec{J} \tag{10}$$

However, it is more useful to calculate the closest approach between the quark line of flight and the track. This can be done by minimizing $\mid \overrightarrow{RQ} \mid^2$, the square distance of two points $\vec{Q}$ and $\vec{R}$ belonging to the quark and particle lines respectively (figure 4). At the minimum, $\vec{Q}$ and $\vec{R}$ are conveniently represented by their abscissa s_J and s_t taken relatively to their origin: $\vec{V}$ for the quark line and $\vec{P_a}$ for the track. When the particle is a b product, the values of s_J and s_t are positive. For that reason, we assign to the track impact parameter δ_a the sign of the s_J abcissa[3]. The expression of s_J is derived in section 6.2.

6.2 Track-jet distance in space

The quantity $\delta_J = \mid \overrightarrow{RQ} \mid$ is the closest approach distance between the quark line and the track. The interest of δ_J is to be only sensitive to cascade decays of the b quark. In the limit of no errors, the quark and the track would intersect exactly when the particle is

[3]We may have taken as well s_t.

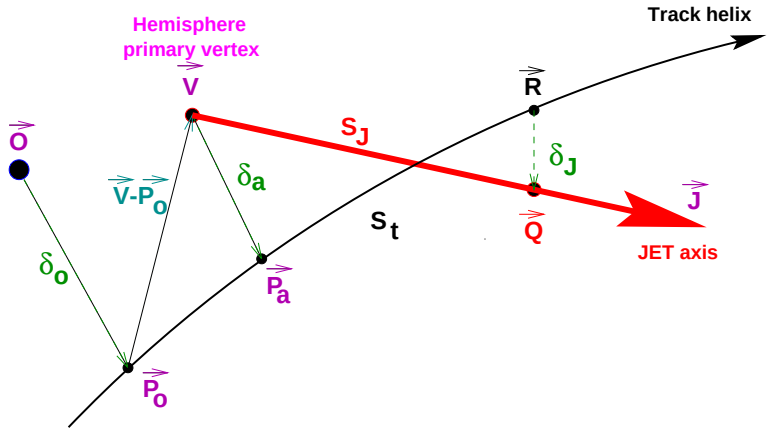


Figure 4: Definition of the signed impact parameter and the track-jet distance.

produced either at the interaction point or at the first generation decay. Therefore, only second generation decays would produce non vanishing values of δ_J.

Mathematically, the problem of finding the point of closest approach between a line and a helix in space is transcendental and an iterative procedure is needed. The procedure has only been applied to tight 3D tracks. For 2D tracks, it is meaningless.

We start by approximating the track as a line defined by the point $\vec{P}_a$ of closest approach of the track to the hemisphere vertex, plus its direction, $\vec{T}_a$. The same for the line of the jet axis, where the origin is the hemisphere primary vertex $\vec{V}$. We compute the arc length s_t along the track which corresponds to the point of closest approach between the linearized track and the jet axis, by

$$s_t = (\vec{V} - \vec{P}_a) \cdot \left[\frac{\vec{T}_a - \vec{J}\left(\vec{T}_a \cdot \vec{J}\right)}{1 - \left(\vec{T}_a \cdot \vec{J}\right)^2} \right] \tag{11}$$

The assumption of the helix to its tangent become not accurate when s_t is not small compared to the radius of curvature. In this case, a new origin $\vec{P}$ of abscissa s_t with tangent $\vec{T}$ has to be substituted to $\vec{P}_a$ and $\vec{T}_a$. The change of origin is explicited in section 6.3 and equation (11) is again solved. The total path from $\vec{P}_a$ is updated and the process is iterated until the path length change is small. This takes generally 1 iteration and a maximum of 4. By following this procedure, the point $\vec{R}$ on the track of closest approach track-jet is obtained as $\vec{R} = \vec{P} + s_t\vec{T}$ with $\vec{P}$, $\vec{T}$ and s_t taken from the last iteration. The corresponding point $\vec{Q}$ on the jet axis is then determined from relation $\vec{Q} = \vec{V} + s_J\vec{J}$, where s_J is defined by

$$s_J = (\vec{V} - \vec{P}_a) \cdot \left[\frac{\vec{T}\left(\vec{T} \cdot \vec{J}\right) - \vec{J}}{1 - \left(\vec{T} \cdot \vec{J}\right)^2} \right] \tag{12}$$

The quantity s_J is just the distance on the jet line between $\vec{V}$ and $\vec{Q}$ (*track-jet abscissa*). The vector $\overrightarrow{RQ}$ can then be written as

$$\overrightarrow{RQ} = \vec{\delta}_a - \left[\frac{\vec{\delta}_a \vec{U}}{|\vec{U}|} \vec{U} + \frac{\vec{\delta}_a \vec{V}}{|\vec{V}|} \vec{V} \right] \tag{13}$$

where $\vec{U} = (\vec{T} + \vec{J})/2$ and $\vec{V} = (\vec{T} - \vec{J})/2$. The *track-jet distance* δ_J is therefore given by the simple formula

$$\delta_J^2 = \delta_a^2 - \frac{\left[\vec{\delta_a} \cdot \vec{U}\right]^2}{|\vec{U}|^2} - \frac{\left[\vec{\delta_a} \cdot \vec{V}\right]^2}{|\vec{V}|^2} \tag{14}$$

Being a minimal distance, $\delta_J < \mid \delta_a \mid$.

6.3 The track helix linearization

For the 3D determination of the impact parameter and the track-jet distance, it is necessary to propagate the track parameters to a new point at the arc length s in space, using a linear approximation of the track.

Given the unitary vector of the tangent $\vec{T}_0 = (T_{x,0}, T_{y,0}, T_{z,0})$ at the point $\vec{P}_0 = (P_{x,0}, P_{y,0}, P_{z,0})$, the tangent parameters $\vec{T}_1$ of the same helix at the arc length in space s are given by the formulas

$$T_{x,1} = T_{x,0} \cos\beta - T_{y,0} \sin\beta \tag{15}$$

$$T_{y,1} = T_{x,0} \sin\beta + T_{y,0} \cos\beta \tag{16}$$

$$T_{z,1} = T_{z,0} \tag{17}$$

where $\beta = s\sqrt{T_{x,0}^2 + T_{y,0}^2}/\rho$ represents the rotation of the helix in the $R\phi$ projection between $\vec{P}_0$ and $\vec{P}_1$. ρ is the projected signed radius. The point $\vec{P}_1$ is defined by

$$P_{x,1} = P_{x,0} + \rho \frac{T_{x,0} \sin\beta - T_{y,0}(1 - \cos\beta)}{\sqrt{T_{x,0}^2 + T_{y,0}^2}} \tag{18}$$

$$P_{y,1} = P_{y,0} + \rho \frac{T_{y,0} \sin\beta + T_{x,0}(1 - \cos\beta)}{\sqrt{T_{x,0}^2 + T_{y,0}^2}} \tag{19}$$

$$P_{z,1} = P_{z,0} + sT_{z,0} \tag{20}$$

6.4 Signed impact parameter in two dimensions

When the experimental track precision in $R\phi$ is much higher than in z (which corresponds to the case when $R\phi$ VD hits have been associated to the track but not in z), a standard 2D impact parameter reconstruction must be adopted (all data taken in 1991, 1992, 1993 and a small fraction of tracks in 1994 and 1995).

Taking as starting point the track parameters at perigee (point of closest approach to the DELPHI origin), the 2D impact parameter with respect to the hemisphere vertex projected on the $R\phi$ plane is

$$\eta_a = h_0^{xy} + (V_y \cos\phi_0 - V_x \sin\phi_0) - \frac{(V_x \cos\phi_0 + V_y \sin\phi_0)^2}{2\rho} \tag{21}$$

where ρ is the signed curvature of the track projected on the $R\phi$ plane. The notation η_a is to avoid confusion with the 3D impact parameter δ_a. The first term of expression (21) corresponds to a coordinate change from the origin of DELPHI to the reconstructed hemisphere primary vertex and the second one is a correction due to the track curvature. Similarly, the impact parameter in z can be estimated according to the expression

$$\Delta_z^a = \Delta_z^0 - V_z + \frac{V_x \cos\phi_0 + V_y \sin\phi_0}{\tan\theta_0} \tag{22}$$

The principle of signing the impact parameters in two dimensions is similar to the case of three dimensions. The impact parameter in $R\phi$ projected on the jet axis can be estimated as

$$q_J = \eta_a \sin\epsilon_J \tag{23}$$

where ϵ_J is the angle (projected on $R\phi$) of the trajectory at perigee with the jet direction. Note that q_J is positive for decay products of B and D hadrons traveling in the downstream direction of the jet.

6.5 Impact parameter errors

The impact parameter is the distance of the trajectory to the primary vertex. The error on this quantity has therefore two components. The first one is due to the track extrapolation error at the DELPHI origin [4]. The second one, which has a smaller contribution, is due to the primary vertex itself. The accuracy on the primary vertex depends on the beam spot size and the accuracy of the tracks included. The error calculation algorithms depend on the kind of track we have, i.e. 2D and 3D metric.

6.5.1 2D impact parameter errors with respect to the hemisphere vertex

At the level of individual tracks, the error on the impact parameters η_a and Δ_z^a are obtained by differentiating equations (21) and (22). The calculation requires to propagate the track impact parameters at perigee (h_0^{xy} and Δ_z^0) to the new reference point, the hemisphere primary vertex $\vec{V}$. As this point is close to the DELPHI origin, the propagation has little effect and equation (21) can be taken at first order. For the $R\phi$ component, the error on h_0^{xy} must be added the contribution due to the error on the xy coordinates of $\vec{V}$:

$$\sigma^2_{\eta_a} = \sigma^2_{h_0^{xy}} + \sin^2\phi_0\sigma^2_{V_x} + \cos^2\phi_0\sigma^2_{V_y} - 2\sin\phi_0\cos\phi_0 cov(V_x, V_y) \tag{24}$$

The z component error $\sigma_{\Delta_z^a}$ is derived from equation (22):

$$\sigma^2_{\Delta_z^a} = \sigma^2_{\Delta_z^0} + \sigma^2_{V_z}(\cos^2\phi_0\sigma^2_{V_x} + \sin^2\phi_0\sigma^2_{V_y})/\tan^2\theta_0$$
$$+ \sin 2\phi_0 cov(V_x, V_y)/\tan^2\theta_0 + 2[\cos\phi_0 cov(V_x, V_z) + \sin\phi_0 cov(V_y, V_z)]/\tan\theta_0 \tag{25}$$

A similar equation is derived for the covariance $cov(\eta_a, \Delta_z^a)$. The correlation due to the fact that the track could be included in the vertex fit is neglected. The error on q_J the impact parameter in $R\phi$ projected on the jet axis is then straightforward. There is an additional error coming from the angular uncertainty on the jet axis direction.

6.5.2 3D impact parameter errors with respect to the hemisphere vertex

One advantage to compute the impact parameter in space instead that in $R\phi$ and z projections separately is that the potential $R\phi - z$ correlation in the track parameters is automatically included in all the physical observables connected with the impact parameters. However, the error of the impact parameter in space is more complicated to estimate. For convenience we express $\vec{\delta}_a$ in a local helix frame in the vicinity of the reconstructed hemisphere primary vertex $\vec{V}$, defined by 3 unitary vectors: $\hat{t}$ and $\hat{n}$ are the tangent and normal (on the $R\phi$ plane) to the track in the $R\phi$ projection and $\hat{k}$ is a vector along the z direction. The vector $\vec{\delta}_a$ can be expressed as a function of η_a and Δ_z^a previously introduced:

$$\vec{\delta}_a = \eta_a \hat{n} + \Delta_z^a \hat{k} \tag{26}$$

It is convenient to define an unit vector $\hat{d}_a$ in the direction of $\vec{\delta}_a$, by $\vec{\delta}_a = \delta_a \hat{d}_a$. For small displacements in the interaction region, the contributions due to errors on track angles can be ignored. The error σ_a on δ_a can be expressed by

$$\sigma_a^2 = \left(\hat{d}_a \cdot \hat{n}\right)^2 \sigma_{\eta_a}^2 + \left(\hat{d}_a \cdot \hat{k}\right)^2 \sigma_{\Delta_z^a}^2 + 2\left(\hat{d}_a \cdot \hat{n}\right)\left(\hat{d}_a \cdot \hat{k}\right) cov(\eta_a, \Delta_z^a) \tag{27}$$

The quantities σ_{η_a}, $\sigma_{\Delta_z^a}$ and $cov(\eta_a, \Delta_z^a)$ are given by equations (24) and (25). The track-vertex correlation effects have been again neglected.

The procedure followed to estimate the error on the track-jet distance δ_J (σ_{δ_J}) is similar to the one described before for the impact parameter in space. The additional component to be considered in the error propagation is the angular uncertainty on the jet axis determination. The jet direction uncertainty can be written as

$$d\vec{J} = d\alpha_J \hat{n}_J + d\beta_J \hat{e}_J \tag{28}$$

where $\hat{n}_J$ and $\hat{e}_J$ are two orthonormal vectors both perpendicular to the jet axis $\vec{J}$. $d\alpha_J$ and $d\beta_J$ represent small displacements along the "north" and "est" directions given by $\hat{n}_J$ and $\hat{e}_J$ respectively. These small displacements are connected to the angular uncertainties in the jet axis measurement. It can approximated that the mean values of both displacements are similar and equal to the jet axis resolution σ_{jet}. In $Z \to b\bar{b}$ events, typical resolutions in the estimate of the B hadron direction of about 70 mrad are obtained, improving to about 50 mrad for jet energies above 10 GeV. The error on δ_J can then be determined applying error propagation to the expression (14). However, a simplest expression for δ_J can be obtained if we take as reference point of the track $\vec{P}_a$ instead of $\vec{P}_0$. In that case $\vec{\delta}_a \cdot \vec{T}_a = 0$ and expression (14) is simplified to

$$\delta_J^2 = \delta_a^2 - \frac{q_J^2}{1 - \left(\vec{T} \cdot \vec{J}\right)^2} \tag{29}$$

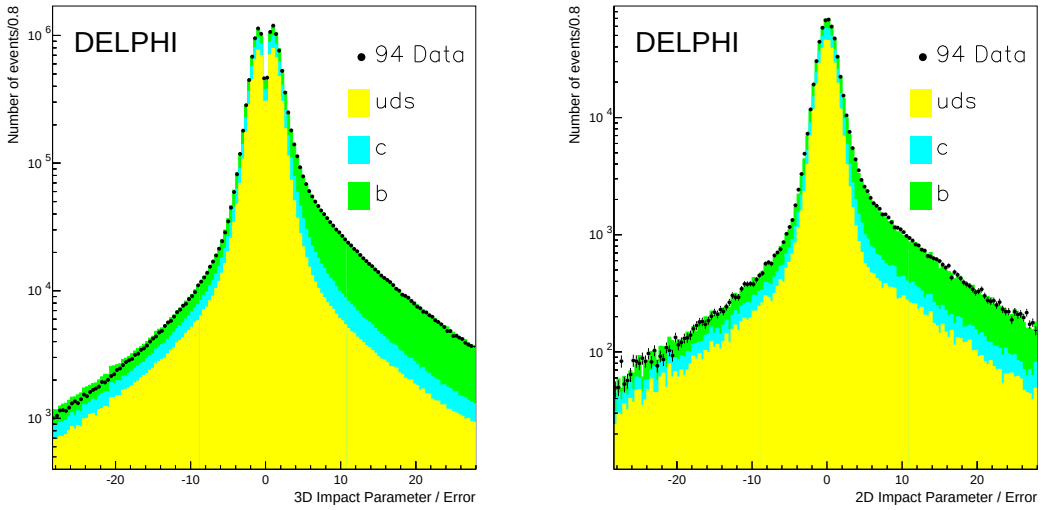
The final expression for σ_{δ_J} can easily be obtained after a little of algebra from equations (10), (26), (28) and (29).

The errors associated to the projected impact parameter on the jet axis q_J (σ_{q_J}) and on the track-jet abscissa s_J (σ_J) are calculated using exactly the same procedure as for the track-jet distance error.

6.6 Impact parameter significance

The ratio between the impact parameter and its error gives the *statistical significance* of the measured impact parameter. Figure 5 represents the significance, $S = \delta_a/\sigma_a$ in 1994 for (a) tight 3D tracks and (b) tight 2D tracks for data and Monte Carlo simulation. For simulation, the composition of uds, c and b quarks is shown. The large positive tail is the lifetime signal. The negative half of the distribution measures the resolution of the impact parameter reconstruction, arising from inaccurate track reconstruction (this sample of tracks is mainly produced at the interaction point and have no true impact parameter). This resolution effect should be equally likely to be positive or negative. In both cases, 3D and 2D metric, the negative part of the resolution is well fitted to the sum of four Gaussians plus one exponential function. These fits are a direct measure of the resolution function $\mathcal{R}(S)$ for the impact parameter significance.

Unfortunately a complete, physically motivated parameterization of the non-Gaussian tail does not exist since there are many sources of completely different nature which produce it. They include unavoidable mistakes in the track search algorithm producing large impact parameters, interactions of the particles with the detector material, decays of long-lived particles (K^0, Λ), presence of secondary vertices, etc. That is why the parameterization is rather complex and arbitrary. The non-Gaussian tail depends significantly on the criteria which are used for the selection of tracks and events.



(a) (b)

Figure 5: Signed impact parameter over the error (significance) with respect the hemisphere vertex in the 1994 period for (a) 3D tight tracks and (b) 2D tight tracks.

7 Tracking tuning

The accuracy of the R_b measurement relies on a close agreement between the observed data distributions and those predicted by the detailed detector simulation. The *physical* events generated [5] are passed through a complex and detailed simulation of the DELPHI detector [6]. In a second step, these simulated raw data are analyzed through the same reconstruction programs [10] as the data. However, after this procedure some disagreements remain between data and simulation in the individual track resolution and in the primary vertex description. It is not drastically large but nevertheless, this can spoil the precise determination of R_b due to data/simulation disagreements.

Both the generation of the intrinsic physical parameters and the simulation of the detector response must be as realistic as possible. In studies of b quark events based on the separation of their origin and decay points, the charged track impact parameter resolution and the primary vertex reconstruction uncertainty are the most crucial part of the detector response. The two most important features to reproduce are the resolution function $\mathcal{R}(S)$ of the impact parameter significance S and the mean number of VD hits associated to tracks. The procedure used for tracking tuning is the DELPHI standard method which is described in reference [11].

8 Probability of primary vertex decay products

The resolution function measured from negative impact parameters tracks can be used to extract the lifetime information from the positive impact parameter tracks by following the method firstly proposed by the ALEPH Collaboration in [12] and adopted by DELPHI[13]. This is done by defining a probability function for the tracks

$$\mathcal{P}_T(S) = \int_{-\infty}^{-|S|} \mathcal{R}(s)ds \tag{30}$$

In order to take into account the number of VD hits, we considered separate resolution functions $\mathcal{R}(s)$ for each configuration. This integrated probability represents the probability that a measurement of the significance S is larger than the observed one. Physically $\mathcal{P}_T(S)$ is interpreted as the probability that the track is consistent with coming from the primary vertex given its measured significance. For a set of N tracks the integrated probability is

$$\mathcal{P}_N = \prod \sum_{j=0}^{N-1} \frac{(-\ln \prod)^j}{j!} \tag{31}$$

By construction, a flat distribution of $\mathcal{P}_N$ is expected for a group of tracks from the primary vertex, provided that significances are not correlated. If the group includes tracks from secondary vertices, the distribution has a peak at low values of $\mathcal{P}_N$. In simulation the distribution of $\mathcal{P}_N$ for light quark is approximately flat, while for b quarks it has a sharp peak at 0. In light quark events there is also a small peak (significantly lower than for b and c quarks) at low values of probability, due to residual tracks from V^0 decays or interactions in the detector material (like e^+e^- pairs). For tagging hemispheres, this variable can be applied to tracks belonging to each hemisphere separately ($\mathcal{P}_H$).

9 Search for secondary vertices

The detection of secondary and tertiary vertices significantly separated from the primary vertex is also a signature of B hadrons. This signature carries some independent information with respect to positive impact parameters, leading to different systematic sensitivity on R_b. We shall call secondary particles produced at the B decay vertex and tertiary particles originating from the charmed hadron which decays later. The two group of particles are disconnected in space, but the low decay multiplicity and short decay ranges together with the limited resolution of the tracking system limit the possibility of separation of the two vertices. Then it happens that decay products are in most of cases merged into a single vertex. Also vertices could be reduced to single tracks.

In order to determine the presence of secondary and even tertiary vertices, a search for disconnected groups (that do not share tracks) of charged particles which intersect in space at a sufficient distance from the primary vertex has been implemented. The search is hierarchical: groups of five or more particles are searched first. If none are found or if so, among particles external to these multiplets, quadruplets are searched. Then the procedure is repeated for triplets, doublets and singlets (group reduced to a single particle).

The intersection $\vec{A}_{sc}$ of the group of tracks is defined from a geometrical fit similar to that of equation (8), but now without the inclusion of the beam spot constraint. The vertex fit probability and the proper decay length of the multiplet is the criteria used to accept the group. The decay length is defined as the distance between the hemisphere primary vertex and the secondary vertex candidate projected on the flight direction $\vec{J}_{sc}$, approximated by the total momentum direction of the multiplet. From the decay length, it is straightforward to compute the proper decay length of the multiplet by the expression $c\tau_0 = c\tau m_{sc}/p_{sc}$, where m_{sc} is the invariant mass of the vertex and p_{sc} its total momentum:

$$c\tau_0 = \overrightarrow{VA_{sc}} \cdot \vec{J}_{sc} m_{sc}/p_{sc} \tag{32}$$

By definition, the distance is signed positive if the range goes in the same direction as the momentum of the multiplet.

Tight tracks involved in the secondary vertex search were required to pass further cuts. They had to have:

- positive impact parameter;
- a momentum p greater than 0.5 GeV/c; and
- a low probability (using equation 8) for the other tracks of the hemisphere to fit a main vertex (less than 1%). This condition is implemented to remove configurations with only one track, which affects essentially the uds flavour. In b hemispheres the multiplicity of secondary tracks is about 5.5 in average and therefore the configuration with a single secondary track is rare. The condition improves slightly the purity of the selection.

Requirements used for the multiplet definition varies with multiplicity, being tighter for triplets and doublets:

- a fit probability $> 10\%$;

- a decay length > 1.0 mm (> 1.5 mm for doublets and triplets);
- a proper decay length > 0.2 mm (> 0.25 mm for doublets and triplets);
- for doublets and triplets, a vertex fit probability for the remaining non associated tracks of the hemisphere $< 10\%$.

For the tracks that have not been associated to none of the previous multiplets, a singlet search is performed at the last stage. Two situations are distinguished. In the first case a multiplet has already been found. There is a good chance for a b hemisphere, where two vertices (one secondary and one tertiary) are in principle present, to have only one charged particle attached to one vertex (this is often the case of a D^+). Then the vertices are not saturated, and information can still be provided by single tracks. The conditions for the search for such singlets are not severe:

- track momentum > 2.0 GeV/c;
- track significance $S > 3.0$;

The second situation is when no multiplets have been found. The configuration is disfavourable because the hemisphere is probably non b. But if it is b, it may happens that both the secondary and the tertiary vertex have only one charged particle attached or seen. For this reason, we look for pairs of singlets, by imposing tighter conditions than previously:

- angle of the track with respect to the most energetic jet of the hemisphere $< 30°$;
- track momentum > 2.0 GeV/c;
- an intersection of two tracks is computed, which allows to compute a proper decay length required to be > 0.20 mm;
- the fit probability of the pseudo-intersection should be greater than 1%, and the probability of the other tracks to be associated in a main vertex $< 1\%$.

As an example, figure 6 shows the distribution of the proper decay length and mass resulting from the search for quintuplets and quadruplets for a 1994 Monte Carlo subsample. For the same data set, table 4 summarizes the performances of the secondary vertex algorithm for each type of configuration. The reached purities of the different configurations are good with a 42.7% of hemispheres having at least one singlet or multiplet found, with a mean purity of 83.0%. For sextuplets, quintuplets and quadruplets having a sizeable total efficiency of about 12%, the purity is higher than 95%. This algorithm will help in tagging performances in the relevant region of high purity for the R_b analysis.

The algorithm described before provides finally a full list of candidate to secondary and tertiary vertices with their proper decay lengths and invariant masses. How these informations are combined to construct tagging variables will be described in section 11.

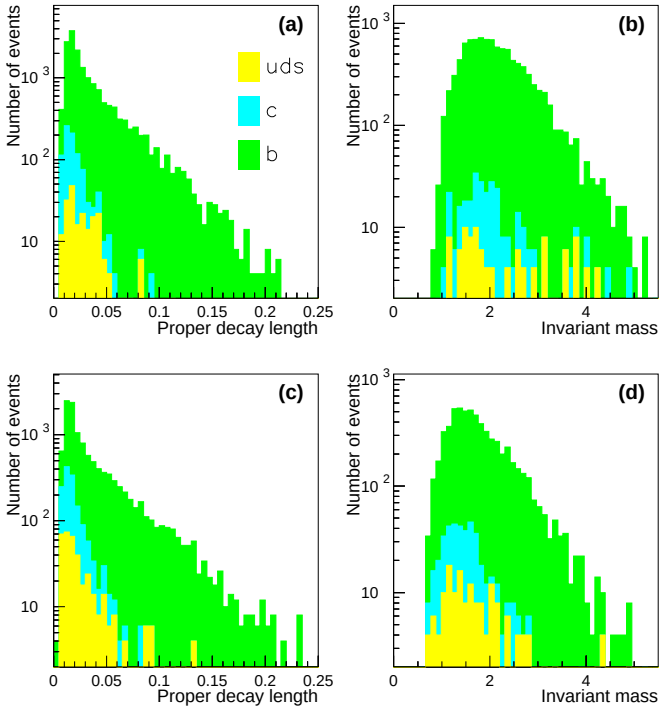


Figure 6: Results of the search for candidates to secondary vertices with five (a,b) and four (c,d) tracks for a 1994 simulation sample. The two most important physical quantities associated to the vertex (proper decay length $c\tau_0$ and invariant mass m_{sc}) are shown. The flavour composition of the selected vertices is also shown. Horizontal scale is in cm. A cut at 0.02 cm is performed on the proper decay length. This cut is already included in the invariant mass distributions.

10 Weights of B hadron decay products

Another information that can be taken from impact parameters is the "counting" of direct secondary particles coming from B hadron decays, prompt as well as cascade. This "counting" can be alternatively done by assigning some kind of probability or weight to each track. In order to optimize the information provided by individual tracks (lifetime as well as event shape properties) several probabilities or weights ω^i are assigned to each particle as a function of:

- the rapidity y of the tight track, defined as

$$y = \frac{1}{2} \ln \left(\frac{E + p_{\parallel}}{E - p_{\parallel}} \right) \tag{33}$$

  where E is the energy of the track and $p_{\parallel}$ its longitudinal momentum with respect to the jet axis;

- the momentum p of the tight track;

Table 4: b efficiencies and purities as a function of several multiplet and singlet configurations found by the secondary vertices search algorithm. These results are obtained from a simulated 1994 data sample.

Hemisphere condition	b purity(%)	b efficiency(%)
None	21.9	100.0
Sextuplets	98.8	3.4
Quintuplets	96.2	4.3
Quadruplets	92.4	4.5
Triplets	86.9	15.2
Doublets	77.9	14.7
Singlets	86.4	26.9
Multiplets	84.7	37.8
Singlets and no multiplets	71.8	4.9
No singlets and multiplets	77.8	15.8
Singlets and multiplets	90.4	22.0
Singlets or multiplets	83.0	42.7

- the impact parameter magnitude over its error, i.e. the significance $S = \delta_a/\sigma_a$ for tight 3D tracks or $S = \eta_a/\sigma_{\eta_a}$ for tight 2D tracks;
- the track-jet abscissa over its error s_J/σ_J for tight 3D tracks and the projected impact parameter on the jet axis over its error q_J/σ_{q_J} for tight 2D tracks;
- the track-jet distance over its error $\delta_J/\sigma_{\delta_J}$ for tight 3D tracks.

The choice of these observables has a direct physical motivation. The rapidity y and the momentum p allows to distinguish between leading and non-leading particles. For instance, momentum and rapidity of D decay products are higher (in average) than in B decays. The significance S and s_J/σ_J (or q_J/σ_{q_J}) are designed to separate tracks originated from non-vanishing lifetime hadrons[4]. Finally the ratio $\delta_J/\sigma_{\delta_J}$ tries to distinguish between prompt secondary tracks and cascade tracks in B decays.

These weights are modelized using the Monte Carlo simulation and they are computed from the ratio of 1D histograms for B decay products over the corresponding 1D histogram for all tracks. In the case of S and s_J/σ_J the weights are computed from 2D histograms in order to include the correlation between both variables. The weights are normalized to its maximum value as it is shown in figure 7 for the 1994/1995 simulation data sample.

From these individual track weights, two global track weights are computed in an attempt to combine the different informations:

$$\mathcal{W}_1^i = \omega^i(y)\ \omega^i(p)\ \omega^i(S, s_J/\sigma_J) \tag{34}$$

$$\mathcal{W}_2^i = \omega^i(y)\ \omega^i(p)\ \omega^i(\delta_J/\sigma_{\delta_J}) \tag{35}$$

$\mathcal{W}_i^1$ and $\mathcal{W}_i^2$ shares the rapidity and momentum dependence, but differs in the lifetime weights. The first one, $\mathcal{W}_i^1$, is sensitive to the impact parameter significance S and the

[4]In the following, the ratios δ_a/σ_a and s_J/σ_J will indicate the proper tight 3D track ratios as well as the corresponding to tight 2D tracks, i.e. η_a/σ_{η_a} and q_J/σ_{q_J} respectively.

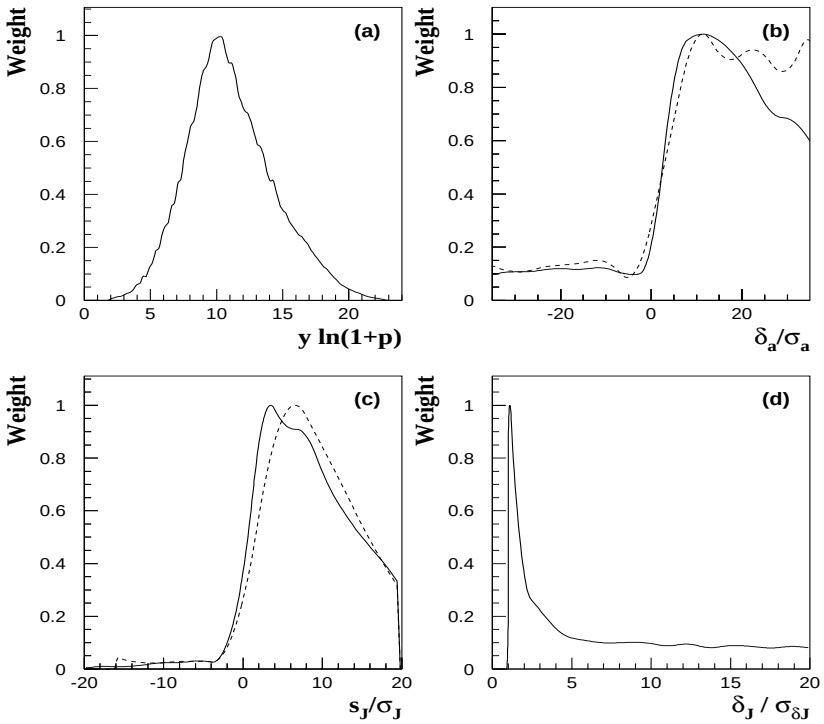


Figure 7: Single track weights of B decay products normalized to its maximum value as a function of: (a) the product of rapidity by the logarithm of the momentum, $y \ln(1+p)$, for all the tight tracks; (b) the significance S for tight 2D (dotted line) and tight 3D (continuous line) tracks; (c) the the track-jet abscissa over its error s_J/σ_J for tight 2D (dotted line) and tight 3D (continuous line) tracks; (d) the track-jet distance over its error $\delta_J/\sigma_{\delta_J}$ for tight 3D tracks.

normalized track jet abscissa s_J/σ_J. The second weight, $\mathcal{W}_i^2$, is sensitive to the track-jet significance $\delta_J/\sigma_{\delta_J}$. There is no strong physical reason for these combinations which may not be optimal, but they seem the most natural ones. How these weights are used in tagging variables is described in section 11.

11 Definition of the multivariate tagging variables

From the ingredients described in previous sections, a set of 13 variables is computed independently in each hemisphere. Some of the variables described in the following were originally proposed in [1]. However, many new variables have been defined and other refined. Here we perform a full description of all the variables.

Figures 8 to 10 display the distributions of these variables for uds, c and b flavours obtained from the simulated sample tuned for the 1994 DELPHI data. Figures are plotted with a logarithmic scale, and the contributions of the three flavours are on top of each other for readability. Real data are superimposed to show the quality of the Monte Carlo description of the data.

11.1 Variables from secondary vertex search

Three variables summarize the results of the secondary vertex search described in section 9. They include multiplicities, masses and proper decay lengths. They are shown in figure 8.

11.1.1 Secondary vertex counter (*SumNSV*)

The *SumNSV* variable tries to count the number of secondary and tertiary tracks. From the number of multiplets and singlets obtained in the secondary vertex search algorithm, it is given by:

$$SumNSV = \sum_{n=1}^{6} nN_n \tag{36}$$

where N_n is the number of multiplets of multiplicity n found.

11.1.2 Secondary vertex proper decay length (*SumDSV*)

The variable *SumDSV* is similar to *SumNSV*. It sums the proper decay lengths of the multiplets weighted by their multiplicities:

$$SumDSV = c\tau_0^0 + \sum_{n=1}^{6} n\overline{c\tau_0^n} \tag{37}$$

where $\overline{c\tau_0^n}$ is the average proper decay length of the multiplets of multiplicity n found in the hemisphere. To the sum is added a default value $c\tau_0^0$. In the case when there is no singlets and multiplets *SumDSV* would be 0. The term $c\tau_0^0$ smears this peak at zero and introduces also some decay length information. $c\tau_0^0$ is a proper decay length computed for all the tracks of the most energetic jet of the hemisphere verifying $p > 1.5$ GeV/c. Apart from this term, when one multiplet is found, *SumDSV* is the product of its proper decay length by the multiplicity.

11.1.3 Secondary vertex mass (*MaxMSV*)

The variable *MaxMSV* is the maximum invariant mass of:

- all the multiplets (multiplicity higher than 1);
- all the possible combinations of pairs formed with all the multiplets and singlets.

Behind considering pairs is the idea that, if secondary and tertiary vertices are separated, they should be combined to make a B hadron.

11.2 Variables using B decay weights

These five variables are weighted counters of B hadron decay products and of some of their characteristics. Figure 9 displays the distribution of these variables for uds, c and b flavours for the 1994 DELPHI data and simulation. The most selective by itself is Ω_1.

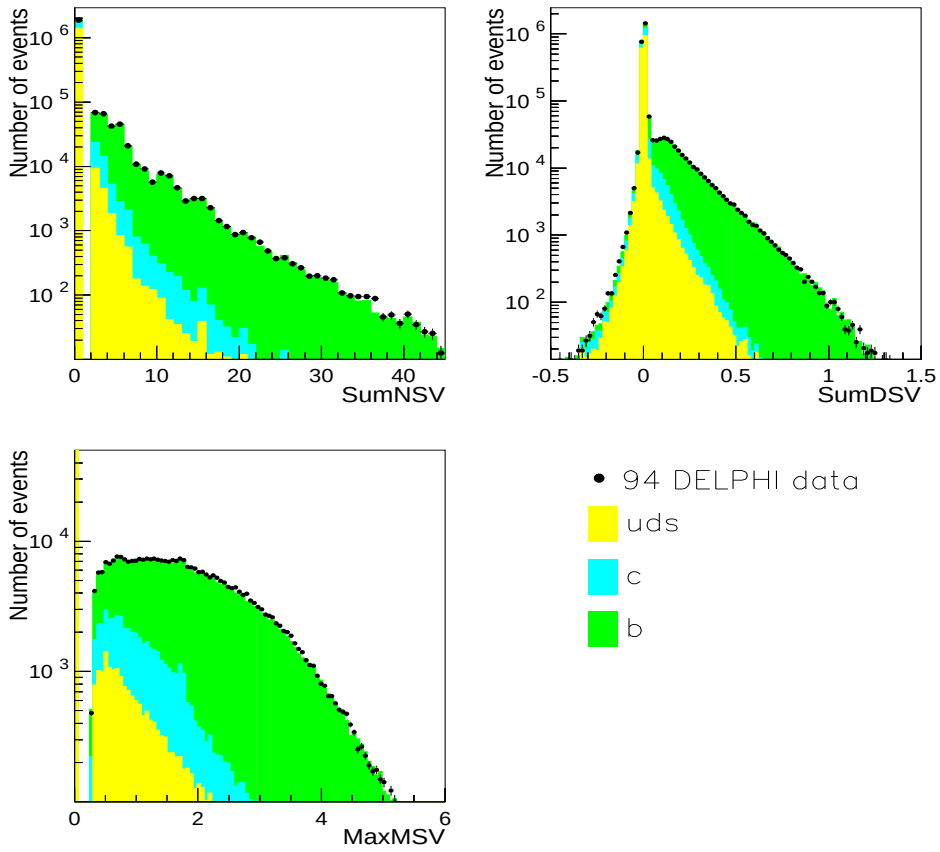


Figure 8: Distribution of b tagging variables from secondary vertex search for the 1994 data sample. Real data are superimposed to show the quality of the Monte Carlo description of the data. For simulation the contribution of uds, c and b flavours is also shown.

11.2.1 Weighted mass (ω_{mass})

This is an adaptation of a variable originally proposed by the ALEPH Collaboration [14]. Particles are first ordered by decreasing consistency to be a B decay product, the criteria being the $\mathcal{W}_1^i$ weight. They are iteratively combined, starting from the track of highest b consistency, until the invariant mass of the group exceeds 2.0 GeV/c. The value of ω_{mass} is defined as the track weight $\mathcal{W}_1^i$ of the last track added. For b hemispheres this can be high since the D hadron mass can be exceeded using only tracks from the B hadron decay; while for c hemispheres ω_{mass} is much smaller, as tracks from the primary vertex are needed to exceed the same cut-off. That mass cut helps in the rejection of c hemispheres in which the D hadron has an unusual long decay length.

11.2.2 Total weight 1 (Ω_1)

The variable Ω_1 is designed to count the total number of secondary particles and is computed as

$$\Omega_1 = \sum_i \mathcal{W}_1^i \tag{38}$$

11.2.3 Total weighted $p_\perp$ ($\Omega_{p_\perp}$)

This variable is defined as the weighted sum

$$\Omega_{p_\perp} = \sum_i \mathcal{W}_1^i p_\perp^2 \tag{39}$$

The sum of $p_\perp^2$ weighted by b probabilities intend to enhance the feature that b products have in average larger $p_\perp^2$. The higher $p_\perp^2$ in b decays is due to the higher mass of the B mesons.

11.2.4 Total weighted p (Ω_p)

This is a weighted variable similar to the previous one, which intends to compute the sum of p of secondary particles:

$$\Omega_p = \sum_i \mathcal{W}_1^i p \tag{40}$$

This sum intends to be large for the b flavour, the B hadron carrying most of the initial quark momentum.

11.2.5 Total weight 2 (Ω_2)

This variable, specific of 3D tracking, is only defined for the 1994 and 1995 data samples. Like Ω_1, the Ω_2 is designed to count the total number of "tertiary" tracks, since the weight $\mathcal{W}_2^i$ based on the track-jet distances:

$$\Omega_2 = \sum_i \mathcal{W}_2^i \tag{41}$$

11.3 Miscellaneous variables

Figure 10 displays the distribution these five variables for uds, c and b flavours for the 1994 DELPHI data and simulation.

11.3.1 Boosted sphericity ($\ln S$)

This variable is the only one computed exclusively with four-momenta. The jet sphericity of the particles belonging to the most energetic jet in the hemisphere is evaluated with respect to an estimated rest frame of a B hadron. The B hadron is assumed to move along the jet direction. A boost, along the jet direction, with a Lorentz γ parameter is needed to perform the transformation from the laboratory frame to the B rest frame. Monte Carlo studies show that at Z energies the optimum value is $\gamma \simeq 4$. The sphericity in this frame is expected to be larger for $b\bar{b}$ events than for the other flavours. The sphericity is defined as [15]

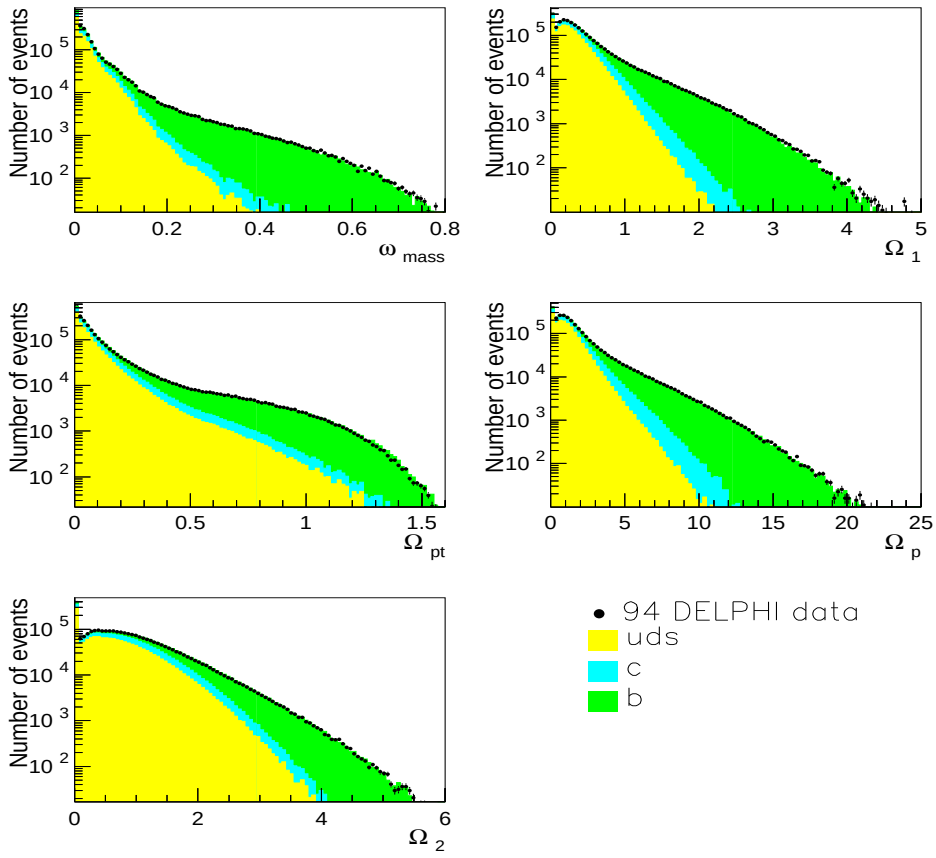


Figure 9: Distribution of b tagging variables from single track B decay weights for the 1994 data sample. Real data are superimposed to show the quality of the Monte Carlo description of the data. For simulation the contribution of uds, c and b flavours is also shown.

$$S = \frac{3 \sum_a \mid \vec{p}_\perp^{\,a} \mid^2}{2 \sum_a \mid \vec{p}^{\,a} \mid^2} \tag{42}$$

where $\vec{p}^{\,a}$ is the three momentum of the a^{th} particle and $\vec{p}_\perp^{\,a}$ is taken relative to the axis which minimizes $\sum_a \mid \vec{p}_\perp^{\,a} \mid^2$ (local sphericity axis).

11.3.2 Normalized decay path (λ)

A "pseudo" secondary vertex fit is attempted in the hemisphere. The most energetic jet of the hemisphere is again associated to the primary quark direction. Only particles making an angle of less than 20° with the jet axis, with an impact parameter with respect to the hemisphere primary vertex $\vec{V}$ of less than 3 mm in space are candidates to the secondary vertex. The fit provides the position $\vec{A}'_{sc}$ of a secondary vertex and its covariance matrix. If there is only one track remaining in the fit, $\vec{A}'_{sc}$ is taken as the intersection in the $R\phi$ projection or in space of this track and the jet axis passing through the main hemisphere vertex $\vec{V}$. If no track is found in the cone, the procedure is applied to the second most energetic jet.

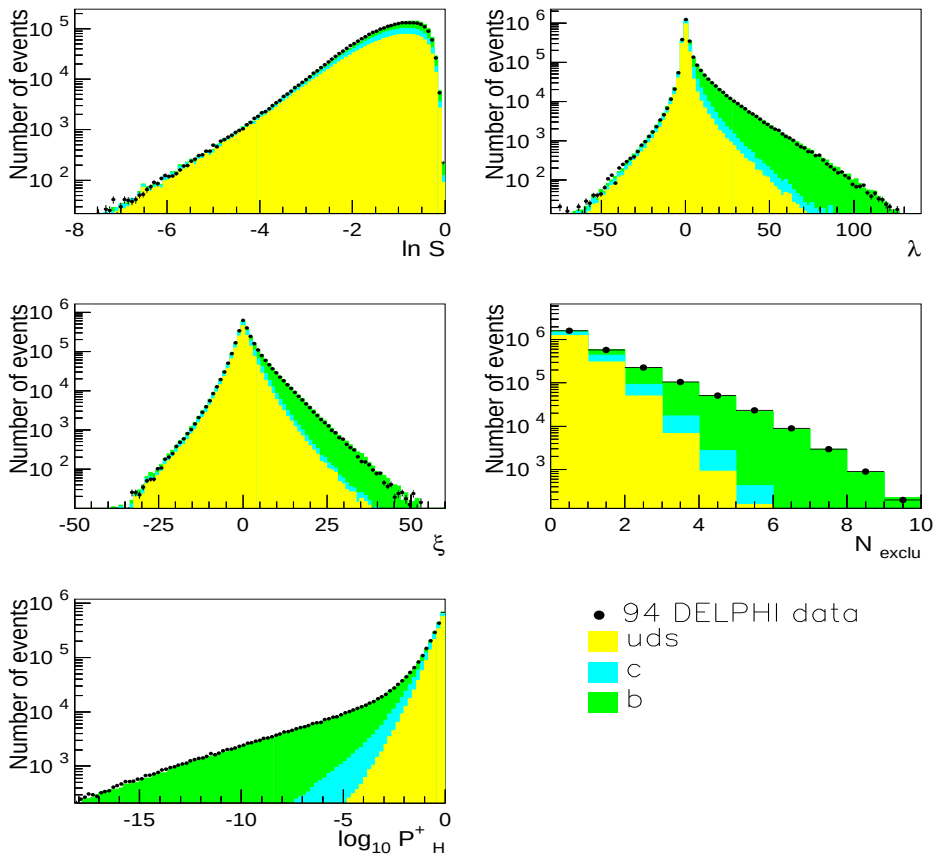


Figure 10: Distribution of several b tagging variables for the 1994 data sample: $\ln S$ is the logarithm of the boosted sphericity of the most energetic jet of the hemisphere, the normalized decay path λ, sum of projected impact parameter ξ, the number of excluded particles N_{exclu} in the primary vertex fit and the hemisphere primary vertex decay products probability (P_H^+). Real data are superimposed to show the quality of the Monte Carlo description of the data. For simulation the contribution of uds, c and b flavours is also shown.

An algebraic distance D along the jet direction $\vec{J}$ is defined for each hemisphere

$$D = \overrightarrow{VA'_{sc}} \cdot \vec{J} \tag{43}$$

and dividing by its error σ_D the "pseudo" normalized decay path variable λ is

$$\lambda = D/\sigma_D \tag{44}$$

11.3.3 Sum of normalized track-jet abscissa (ξ)

The sum of the normalized track-jet abscissa is defined for tight 3D tracks as

$$\xi = \sum_i s_J^i/\sigma_J^i \tag{45}$$

and for tight 2D tracks it is replaced by the normalized projected impact parameter:

$$\xi = \sum_i q_J^i/\sigma_{q_J^i} \tag{46}$$

The ξ distribution is expected to be centered at zero for the uds flavour while for c and b an asymmetry in the positive direction is expected, due to the fact that decay products have positive track-jet abscissa or projected impact parameter.

11.3.4 Excluded particles (N_{exclu})

N_{exclu} is the number of excluded particles during the iterative procedure of the hemisphere vertex fit described in section 5. This variable, which is correlated to the weighted sum Ω_1, is highly selective by itself.

11.3.5 Hemisphere primary vertex probability ($\mathcal{P}_H^+$)

This variable is basically the standard DELPHI b tagging criteria, whose definition has been detailed in section (8). Originally proposed by ALEPH [12], this probability was adapted to DELPHI on the basis of a common event vertex [13]. However, in this analysis, the recalculation of a primary vertex distinct for each hemisphere imposes to recompute the variable in order to redefine the significance S and the resolution function $\mathcal{R}(S)$. The analytical parameterization of the resolution function was computed separately for tight 2D and 3D tracks. As tight 2D and 3D tracks may be found together in the same hemisphere, the individual track probabilities take into account the type of each track, and the calculation of the global probability $\mathcal{P}_N$ given in equation (31) can be done. To increase the selection power of the variable, only tracks with positive impact parameter (which contain the lifetime information) are included in $\mathcal{P}_N$.

12 Flavour confidences

In order to improve the performances of the multivariate technique, we have tried to incorporate the know-how of other multivariate-like techniques developed by DELPHI into a global multivariate classifier. Such a very interesting and elaborated technique, called *flavour confidences*, was proposed in reference [16]. Similarly to the multivariate approach,

the confidence method is based not only on the track impact parameters but also on two other kinematic variables, the track momentum and the angle with respect to the jet axis. The track information is used differently in both techniques, so the overlap between them is expected not to be complete and interesting gains in performances can be obtained in a combination. Like the variable $\mathcal{P}_H^+$ described in 11.3.5, these confidences has been adapted with respect to [16] to the reconstruction of one primary vertex for each hemisphere.

The method uses the simulation to build a function which gives the fraction of tracks which come from b, c and uds quarks in a bin of the three particle characteristics: impact parameter over its error δ_a/σ_a, momentum p and angle ϕ to the jet axis. There are kinematic effects in the decay of B hadrons which produce correlations between the three quantities, but they are automatically taken into account by the 3D binning.

For each single track an individual flavour confidence is computed as

$$\mathcal{C}_l(\delta_a/\sigma_a, p, \phi) = \frac{f_l(\delta_a/\sigma_a, p, \phi)}{f_{uds}(\delta_a/\sigma_a, p, \phi) + f_c(\delta_a/\sigma_a, p, \phi) + f_b(\delta_a/\sigma_a, p, \phi)} \tag{47}$$

where

$$f_l(\delta_a/\sigma_a, p, \phi) = \frac{N_l(\delta_a/\sigma_a, p, \phi)}{N_l^{total}} \tag{48}$$

$N_l(\delta_a/\sigma_a, p, \phi)$ is the number of events in the bin $(\delta_a/\sigma_a, p, \phi)$ with initial quark flavour l (taken from simulation) and N_l^{total} is the total number summed over all bins. The individual flavour confidences must be combined to make the hemisphere tag:

$$\mathcal{CONF}_{uds} = \frac{3\prod_i \mathcal{C}_{uds}^i}{3\prod_i \mathcal{C}_{uds}^i + \prod_i \mathcal{C}_c^i + \prod_i \mathcal{C}_b^i} \tag{49}$$

$$\mathcal{CONF}_c = \frac{\prod_i \mathcal{C}_c^i}{3\prod_i \mathcal{C}_{uds}^i + \prod_i \mathcal{C}_c^i + \prod_i \mathcal{C}_b^i} \tag{50}$$

$$\mathcal{CONF}_b = \frac{\prod_i \mathcal{C}_b^i}{3\prod_i \mathcal{C}_{uds}^i + \prod_i \mathcal{C}_c^i + \prod_i \mathcal{C}_b^i} \tag{51}$$

$\mathcal{C}_l^i$ being the l-flavour confidence for track i. The factor 3 has the same physical motivation as in equations (2), (3) and (4). Figure 11 displays the distribution the hemisphere confidences for uds, c and b flavours for the 1994 DELPHI data and simulation.

13 Combined multivariate/confidences tagging

The two tags, multivariate and confidences, can be combined using a simple linear combination for each flavour. In order to be homogeneous with the multivariate flavour likelihoods $\mathcal{L}_{uds}$, $\mathcal{L}_c$ and $\mathcal{L}_b$, we have to take the logarithm of the difference to unity of each flavour confidence:

$$\Delta_{uds} = (1-\alpha)\mathcal{L}_{uds} - \alpha \ln(1 - \mathcal{CONF}_{uds}) \tag{52}$$

$$\Delta_c = (1-\alpha)\mathcal{L}_c - \alpha \ln(1 - \mathcal{CONF}_c) \tag{53}$$

$$\Delta_b = (1-\alpha)\mathcal{L}_b - \alpha \ln(1 - \mathcal{CONF}_b) \tag{54}$$

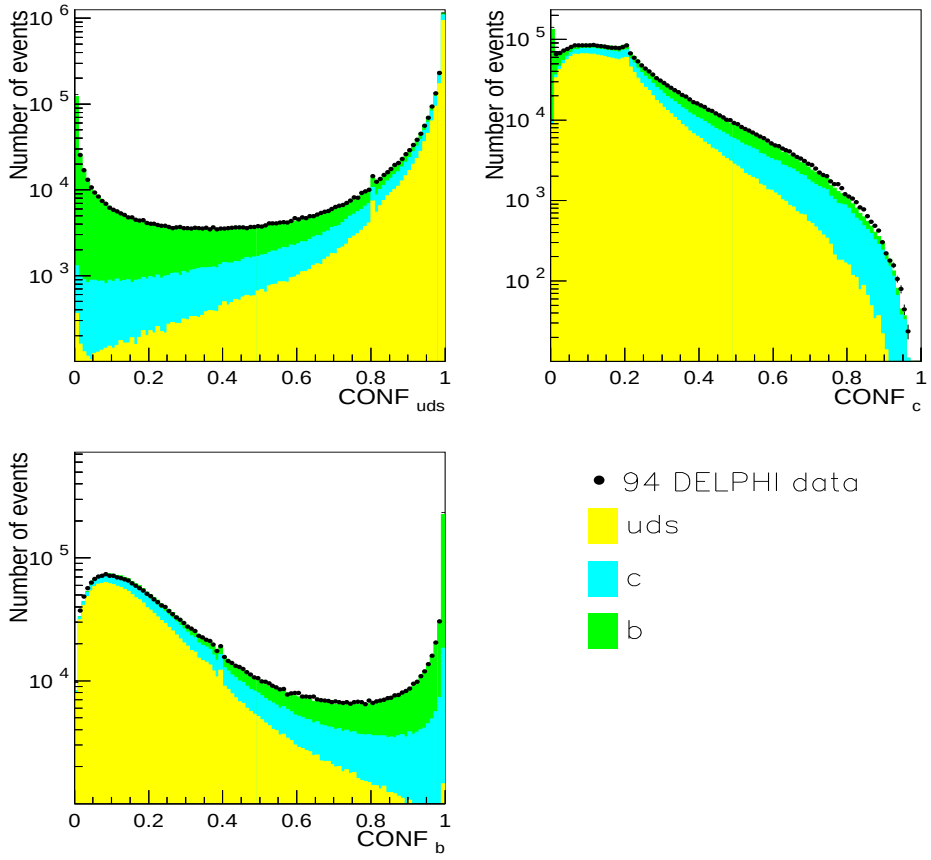


Figure 11: Distribution of uds,c and b confidences in 1994 simulation and data. Real data are superimposed to show the quality of the Monte Carlo description of the data. For simulation the contribution of uds, c and b flavours is also shown.

The quantities Δ_{uds}, Δ_c and Δ_b are called *flavour multivariate discriminators* and are the basis of the classification. This way to combine has been proven to be the best of several tried.It could also be possible to optimize one value of α different for each flavour, but it happens that in practice the same value optimize the three flavours. The quoted value was $\alpha = 0.8$. $\mathcal{L}_{uds}$, $\mathcal{L}_c$ and $\mathcal{L}_b$ are calculated from the 13 variables described in section 11, while the flavour confidences are based on the 3 variables of section 12. The apparently high ratio $\alpha/(1-\alpha) = 4$ is due to the fact that the range definition of the multivariate flavour likelihoods is higher than the corresponding to the flavour confidences. It corresponds approximatively to an equal weight of the two components. Figure 12 show the distributions of the flavour multivariate discriminators for all the 1991 to 1993 and 1994+1995 data and simulation separately. It can be seen that the agreement between data and Monte Carlo is good, thanks to a very very fine physics and detector tuning of the simulation.

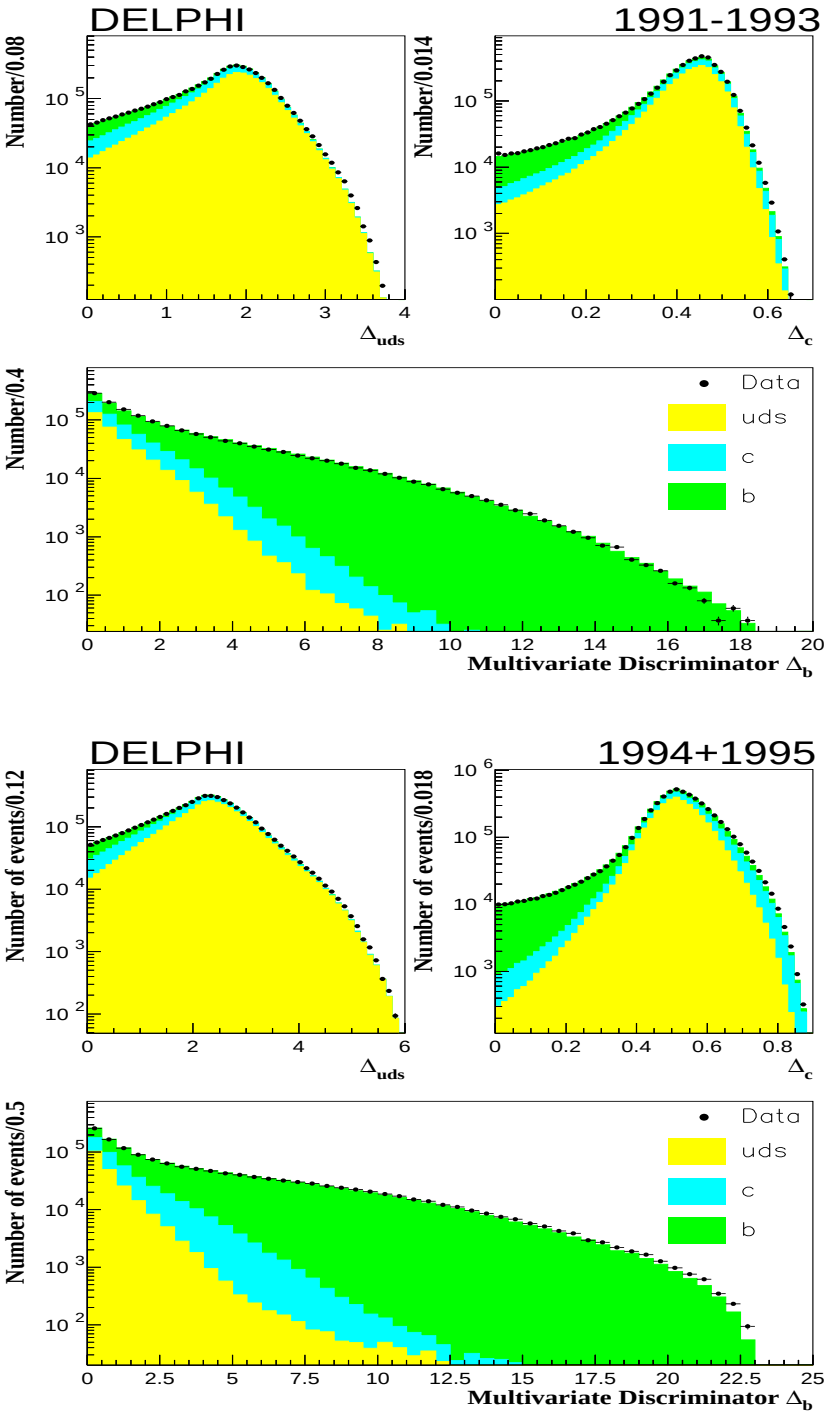


Figure 12: Distribution of the multivariate discriminator Δ in the uds, c and b tags for all the 1991 to 1993 (above) and 1994+1995 (below) data and simulation. The different types of shading show the different flavour contributions to the simulated event sample. The simulation distributions are normalized to the data statistics.

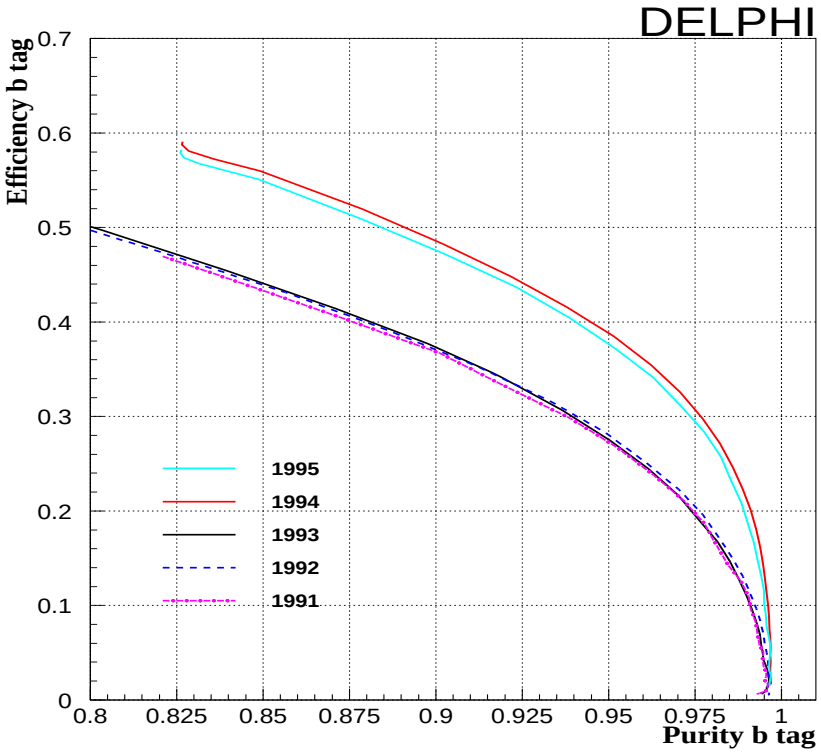


Figure 13: The hemisphere b efficiency obtained as a function of the b purity in tagging hemispheres with the multivariate technique for each year.

14 Tagging performances

The efficiency of the hemisphere b tag as a function of the b purity for each data set is given in figure 13. Figure 14 plots for the three tags the *background efficiencies* versus the tag efficiency. The background efficiencies are the probability to classify the wrong flavours in a given tag. Results have been averaged and presented separately for the 1991-1993 and 1994+1995 periods, since the different microvertex setup leads to very different tagging performances. The plots are obtained for hemispheres within an angular acceptance of 0.65 on $\mid \cos\theta_{thrust} \mid$. From figure 13, for purities of 90%, the efficiency is approximately 48% in 1994+1995 and about 37% in 1991-1993. At 95% purity, efficiencies are about 38% and 28% respectively. At 98% purity, the efficiencies drop to about 28% and 18%. Reading figure 14, for a b efficiency of 20%, the mistag probabilities are: a) in 1994+1995, less than 0.02% for uds quarks and 0.2% for c quarks, and b) in 1991-1993, about 0.04% for uds and 0.6% for c quarks. Therefore very high purities can be reached in the b identification with certainly important b efficiencies.

It should be stressed that this tool provides also uds and c tags. Their performances are by far poorer than of the b tag. For instance, for a 15% uds tag efficiency, the background efficiencies are about 5% for c quarks and less than 1% for b quarks, for all data. For a 15% c tag efficiency, the background efficiencies are less than 5% for both uds and b quarks in 1994+1995. In 1991-1993, for the same efficiency, the uds background is about 7% and the b background less than 7%. Figure 15 shows the efficiencies of the hemisphere uds and

c tags as a function of the corresponding purities for each data sample. Interesting is the improvement in c performances of the 1994+1995 data sample with respect to the 1991-1993. These tags can be used alone or combined between them and with the powerful b tag. For example, the quality of the uds and c tags can be improved by reducing b contamination by imposing extra cuts on the b multivariate discriminator $\mathcal{L}_b$.

However, although the uds and the c tags are poor compared to the b tag ones, both tags can help in the rejection of b tag backgrounds for the precise R_b determination. Moreover, and what it is more interesting, they are a fundamental part of the technique used to self-calibrate the tagging, so reducing dependences on simulation models and therefore reducing important systematic uncertainties affecting the R_b determination.

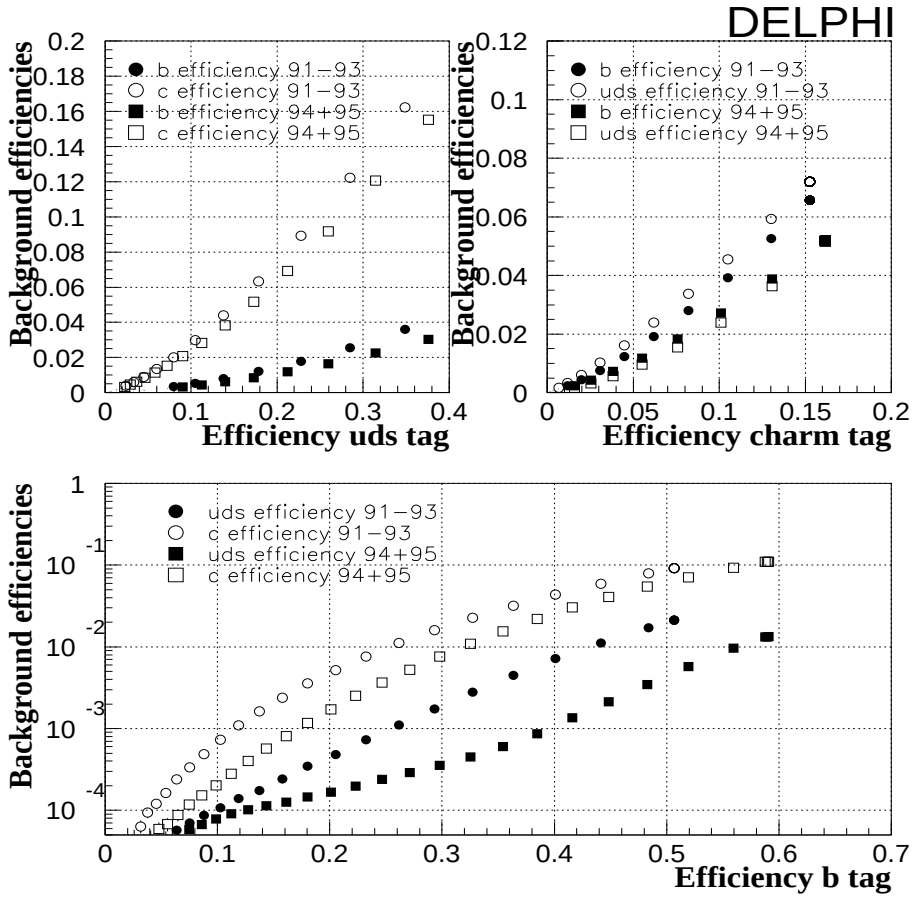
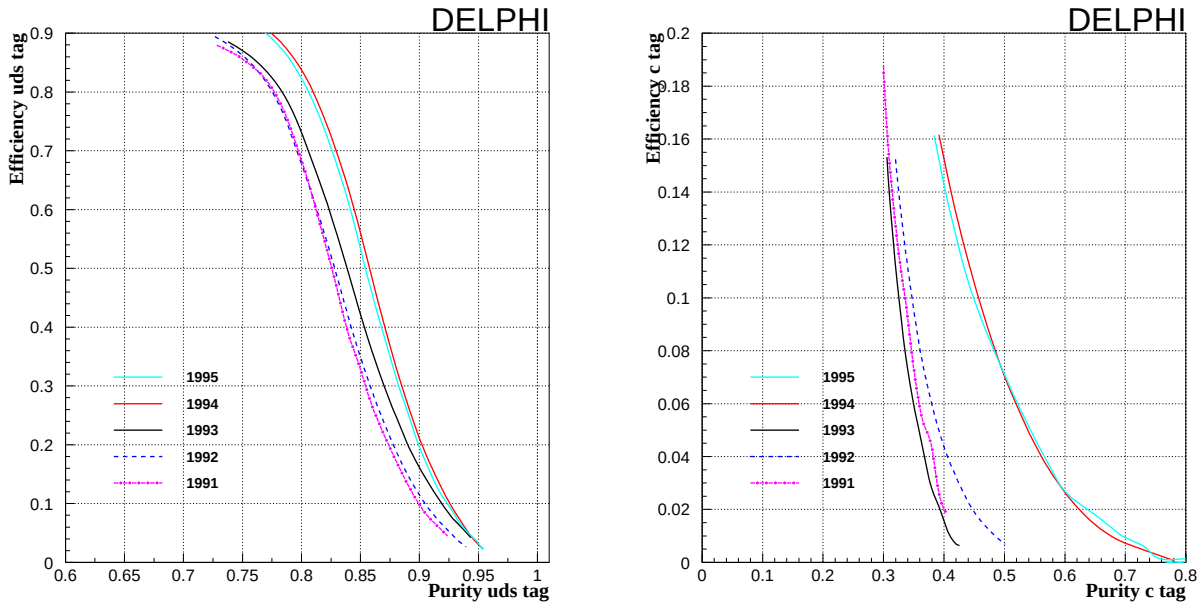


Figure 14: The hemisphere backgrounds in each flavour tag as a function of the corresponding flavour efficiency with the multivariate technique. The quoted performances are shown separately for 1991-1993 data and 1994+1995 data separately due to the different tagging performances due to the different microvertex detector setup.

15 Hemisphere-hemisphere tagging correlation

The hemisphere-hemisphere tagging correlation coefficient is defined as



(a) (b)

Figure 15: The hemisphere uds and c efficiency obtained as a function of the uds and c purity in tagging hemispheres with the multivariate technique for each year.

$$\rho_b = \frac{double\ tag\ rate(\Delta_b^{Hem\ 1}, \Delta_b^{Hem\ 2})}{single\ tag\ rate(\Delta_b^{Hem\ 1})\ single\ tag\ rate(\Delta_b^{Hem\ 2})} - 1 \tag{55}$$

where $\Delta_b^{Hem\ 1}$ and $\Delta_b^{Hem\ 2}$ are the multivariate discriminator cuts for hemispheres 1 and 2 respectively. Figure 16 shows the correlation coefficient when cutting on the same value on both sides of the event. By following the method described in [2, 3] to split the total correlation into sources, it can be shown that this correlation is basically due to the physics from gluon radiation. All correlation sources described in section 5 all therefore practically eliminated. Other smaller sources of correlation are due to angular effects of the vertex detector acceptance.

16 Giving access to the multitag

The multivariate tags have been stored on *ntuples* [17], produced for real data samples and the standard simulation samples. For an event, it can be accessed easily with the help of two subroutines. To open a new tagging file, the subroutine

```
CALL MDADEB(LUNTU,NTUNAM,IFLAG)
```

has to be called with the parameters:

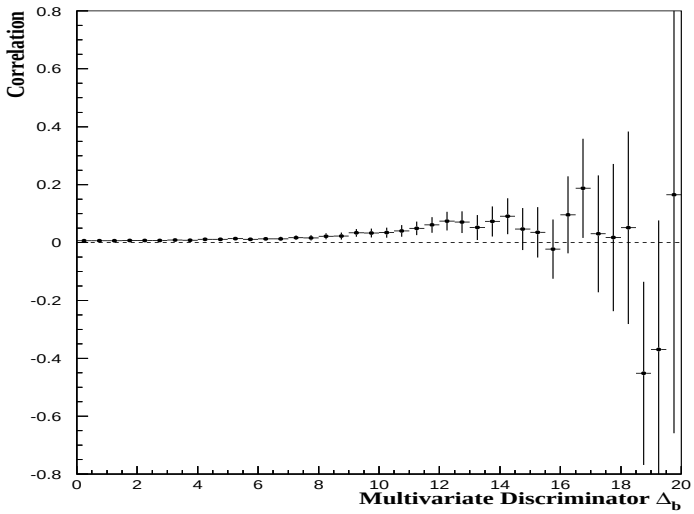


Figure 16: Double b tag correlation coefficient. Points are bin to bin correlated.

I `LUNTU` number of logical unit to assign to the file;

I `NTUNAM` full name (including the path) of the file containing the *ntuple* of tags;

O `IFLAG` error code, 0 if ok.

Then, by calling the subroutine

```
CALL MDATAG(NRUN,NEVT,COSPHECUT,IFLA,NATQ,EVID,DMULT1,DMULT2)
```

once by event, the hemisphere tagging information is provided. Here the parameters are the following:

I `NRUN` run number of the event;

I `NEVT` event number;

I `COSPHECUT` $|\cos\theta_{thurst}|$ cut for angular acceptance;

O `IFLA` output code:

- **0** run and event numbers matched, and inside acceptance;
- **-1** run and event numbers matched, but outside acceptance;
- **-2** run number matched, but event number not matched;
- **-3** run number not matched;
- **-4** disagreement for the `NRUN` address;

O `NATQ` flavour of the event (1=uds, 2=c, 3=b or 0 if unknown in simulation. Always 0 in data);

O `EVID` event and hemisphere identifier (4 words):

EVID(1) run number;

EVID(2) event number;

EVID(3) $\cos\theta_{thurst}$;

EVID(4) ϕ_{thurst};

O `DMULT1` multivariate discriminators Δ_{uds}, Δ_c, Δ_b for hemisphere 1:

DMULT1(1) uds flavour;

DMULT1(2) c flavour;

DMULT1(3) b flavour;

O `DMULT2` multivariate discriminators Δ_{uds}, Δ_c, Δ_b for hemisphere 2:

DMULT2(1) uds flavour;

DMULT2(2) c flavour;

DMULT2(3) b flavour;

$\cos\theta_{thurst}$ and ϕ_{thurst} define the direction of the event thrust axis for hemisphere labelled as 1. If anyone wishes to have access to the multivariate tags, please contact the authors (`martinef@afsmail.cern.ch`) for assistance and to provide the subroutines and the tagging files.

17 Conclusions

The algorithm and performances of a multivariate flavour tagging algorithm at LEP 1 have been described with detail for the full LEP 1 data taken by DELPHI. Interesting improvements in b and c tagging performances are quoted in the 1994/1995 periods with respect to the previous years due to the use of precise three dimensional information and improved tracking capabilities of the detector. Tagging variables incorporate two and three dimensional impact parameter reconstruction, transverse momenta, secondary vertices and invariant mass reconstruction. The tagging technique has been optimized for the precise analysis of R_b, in particular what refers to the separated reconstruction of the primary vertex for each hemisphere in order to achieve an almost hemisphere tagging independence. However, the flavour tagging can be used in other DELPHI analyses without any significant modifications.

Finally, it should be underlined that the measurement of R_b provides also an accurate measurement of the tagging probabilities. The multivariate tag can be calibrated on the data itself. This allows to reduce the dependence on simulation if one wants to separate the contribution of uds, c and b flavours.

Acknowledgements

This work was facilitated by a collaboration between IFIC in Valencia and LPNHE in Paris, which was funded by a general agreement between CICYT in Spain and the IN2P3 in France.

References

[1] P. Billoir et al., Nucl. Inst. Meth. **A360** (1995) 532.

[2] DELPHI Collaboration, P.Abreu et al., Z. Phys. **C65** (1995) 555.

[3] DELPHI Collaboration, P.Abreu et al. Z. Phys. **C66** (1995) 323.

[4] P. Abreu et al. (DELPHI Collaboration), Nucl. Instr. and Meth. **A378** (1996) 57.

[5] JETSET Monte Carlo program : T. Sjöstrand, Comp. Phys. Comm. **27** (1982) 243; **28** (1983) 229; **39** (1986) 347; T. Sjöstrand and M. Bengtsson, Comp. Phys. Comm. **43** (1987) 367. **B298** (1993) 236.

[6] DELSIM Reference Manual, DELPHI 87-98 PROG 100.

[7] P. Abreu et al. (DELPHI Collaboration), Zeit. Phys. **C73** (1996) 11.

[8] C. Peterson, D. Schlatter, I. Schmitt and P. Zerwas, Phys. Rev. **D27** (1983) 105.

[9] P. Billoir et al., Nucl. Inst. Meth. **A241** (1985) 115.

[10] *DELANA User's Guide*, DELPHI 89-44 PROG 137.

[11] G. Borisov and C. Mariotti, Nucl. Inst. Meth. **A372** (1996) 181.

[12] D. Brown and M. Frank, *Tagging b hadrons using track impact parameters*, ALEPH 92-135 PHYSIC 92-124.

[13] G. Borisov, *Lifetime tag of events $Z \to b\bar{b}$ with the DELPHI detector. AABTAG program*, DELPHI 94-125 PROG 208.

[14] R. Barate at al., (ALEPH Collaboration), *A measurement of R_b using a lifetime-mass tag*, CERN-PPE/97-017. To be appear in Phys. Lett. **B**.

[15] Ed. G. Altarelli, R. Kleiss and C. Verzegnassi, *Z physics at LEP*, Vol.1, CERN 89-08, Geneva, 1989.

[16] W.J.Murray, *Improved B tagging using impact parameters*, DELPHI 95-167 PHYS 581.

[17] CERN Program Library, *HBOOK Reference Manual.*