UNIVERSITE BLAISE PASCAL IN2P3

LABORATOIRE DE PHYSIQUE CORPUSCULAIRE

63177 AUBIERE CEDEX

TELEPHONE : 73 40 72 80
TELECOPIE : 73 26 45 98

# MULTIVARIATE ANALYSIS METHODS TO TAG b QUARK EVENTS AT LEP/SLC

B. Brandl[+], A. Falvard[++], C. Guicheney[++],
P. Henrard[++], J. Jousset[++], J. Proriol[++]

[+] Institut für Hochenergiephysik
University of Heidelberg
D-6900 HEIDELBERG GERMANY

[++] Laboratoire de Physique Corpusculaire
de Clermont-Ferrand
CNRS/IN2P3 - Université Blaise Pascal
F-63177 AUBIERE CEDEX FRANCE

*To Appear in Nuclear Instruments and Methods*

PCCF RI 9202

PCCF-RI 92-02

# Multivariate Analysis Methods to Tag $b$ Quark Events at LEP/SLC

B. Brandl[+], A. Falvard[++], C. Guicheney[++],
P. Henrard[++], J. Jousset[++], J. Proriol[++]

[+] Institut für Hochenergiephysik
University of Heidelberg
D–6900 HEIDELBERG GERMANY

[++] Laboratoire de Physique Corpusculaire
de Clermont–Ferrand IN2P3 - CNRS
Université Blaise Pascal
F–63177 AUBIERE CEDEX FRANCE

## Abstract

Multivariate analyses are applied to tag $Z \to b\bar{b}$ events at LEP/SLC. They are based on the specific $b$-event shape caused by the large $b$-quark mass. Discriminant analyses, classification trees and neural networks are presented and their performances are compared. It is shown that the neural network approach, due to its non-linearity, copes best with the complexity of the problem. As an example for an application of the developed methods the measurement of $\Gamma(Z \to b\bar{b})$ is discussed. The usefulness of methods based on the global event shape is limited by the uncertainties introduced by the necessity of event simulation. As solution we present a double tag method which can be applied to many tasks of LEP/SLC heavy flavour physics.

*To Appear In Nuclear Instruments and Methods*

# 1 Introduction

One of the main parts of the physics program at LEP is the precise test of the Electroweak Standard Model [1]. In this respect the $b$-quark sector offers specific aspects of particular importance through the $b$-asymmetry and the partial width $\Gamma(Z \rightarrow b\bar{b})$ measurements. Other important prospects, like the search for $Z \rightarrow \gamma H^0 \rightarrow \gamma b\bar{b}$, can only be achieved if efficient $b$-tagging methods are available. 2 millions of $Z$ decays have been collected by the four LEP experiments in the last two years, and in near future LEP will be operated in Pretzel-scheme with more than four bunches. This will significantly increase the amount of data. To achieve these ambitious physics aims it is mandatory to improve the $b$-tagging.

Compared to other quark flavours $b$-quarks have a larger mass and a longer mean lifetime. While the lifetime information can be summarized essentially by one variable, for instance the distance between the primary and secondary vertex, the mass information is more diluted in the event. To take advantage of the various sources of information, multivariate analyses can be used. It must be understood that ultimately none of the two tags by either mass or lifetime are sufficient when used alone. This is evident since the $D^+$-mesons have mean lifetimes approximately identical to that of $B$-hadrons. This paper stresses the discrimination by the mass difference which requires specific multivariate treatment. Any additional information relevant for $b$-tagging, and especially the lifetime information, can be included in the methods without specific developments.

First used in an industrial framework [2], multivariate analyses have been used in the last few years in high energy physics [3, 4]. They can now be considered as an established tool of heavy quark physics at LEP. Results have been recently published by the ALEPH collaboration using methods described hereafter [5].

Multivariate analyses can be divided in linear methods, like discriminant analyses, and non-linear methods, like neural networks. In the paper we describe the most popular methods and compare their performances with respect to the $b$-quark tagging. Moreover as an example of a specific application, the measurement of $\Gamma(Z \rightarrow b\bar{b})$ at LEP/SLC using these methods will be discussed.

# 2 Methods

This section describes several methods which are applied to tag $e^+e^- \rightarrow b\bar{b}$ events at the $Z$ pole. The techniques are based on pattern recognition. Starting from a fixed number of variables, which have been selected according to the characteristics of the events to be recognized, a classification tool called classifier, is derived .

The derivation is divided into two steps:

- A so-called supervised learning step, in which the system is "taught" which class an event belongs to. This step is commonly performed with simulated events.

- A second step called validation, in which one checks whether the system is able to classify unknown events correctly.

We will discuss linear methods (discriminant analyses, classification trees) and non-linear methods (neural networks).

Throughcut this paper an event $k$ described by $p$ variables is refered to as a vector with $p$-components:

$$\vec{e}_k = (e_{k1}\,,\, e_{k2}, \cdots, e_{kp}).$$

## 2.1 Linear Discriminant Analyses

Discriminant analyses assume a multidimensional normal distribution of the $p$ variables characterized by mean values $\vec{\mu}$=$(\mu_1, ..., \mu_p)$ and a common covariance matrix $\Sigma$ of the different classes. However one can prove that the classification rules are strong enough to be applied on non-gaussian variables too.

Linear discriminant analysis methods derive a linear combination of the selected variables, which provides the best characterization of the difference between the classes. The linear combination is the classifier.

### 2.1.1 Fishers' Linear Discriminant Analysis (FLDA)

Considering two arbitrary classes $C_i$, $C_j$ with mean values $\vec{\mu}_i$, $\vec{\mu}_j$ and assuming the $\Sigma$ matrix to be the same for both classes, the learning step consists in computing $\Sigma$, $\vec{\mu}_i$ and $\vec{\mu}_j$. Classifying an event then means to compare the value of Fishers' discriminant function $L$ [6] for each event $\vec{e}_k$

$$L(\vec{e}_k) = (\vec{\mu}_i - \vec{\mu}_j)[\Sigma^{-1}](\vec{e}_k)^T$$

with

$$\frac{1}{2}\,(\vec{\mu}_i - \vec{\mu}_j)[\Sigma^{-1}](\vec{\mu}_i + \vec{\mu}_j)^T.$$

The linear discriminant function maximizes the norm of the vector $\vec{\mu}_i - \vec{\mu}_j$, which gives the distance of the two classes, and $\vec{e}_k$ will be classified inside the class $C_i$ if

$$L(\vec{e}_k) \geq \frac{1}{2}\,(\vec{\mu}_i - \vec{\mu}_j)[\Sigma^{-1}](\vec{\mu}_i + \vec{\mu}_j)^T$$

and inside the class $C_j$ if:

$$L(\vec{e}_k) \leq \frac{1}{2}\,(\vec{\mu}_i - \vec{\mu}_j)[\Sigma^{-1}](\vec{\mu}_i + \vec{\mu}_j)^T.$$

One can show that such a rule of classification minimizes Baye's risk of misclassification [7].

### 2.1.2 Canonical Discriminant Analysis (CDA)

CDA derives the linear combination of the $p$ selected variables that has the highest possible multiple correlation with the $q$ event classes. Though FLDA and CDA combine the variables linearly they are based on different methods of statistical mathematics. CDA is equivalent to canonical correlation analysis [8]. A rigorous derivation of CDA can be found elsewhere [9]. The partial differentiations involved are complicated, because two sets of weights have to be fitted simultaneously. Here we will only describe how the coefficients of the linear discrimination model are derived.

Describing event $k$ by $\vec{e}_k$ and a dummy vector variable $\vec{h}_k$, which defines the class the event belongs to, then the conditions

$$u_k = \vec{c} \cdot \vec{e}_k$$

$$v_k = \vec{d} \cdot \vec{h}_k$$

$$R_c = \left( \frac{1}{n} \sum_{k=1}^{n} u_k v_k \right) |_{max}$$

where $n$ is the total number of events, define the linear coefficients $c_1, c_2, ..., c_p$ of the canonical discrimination model.

$R_c$ is called "first canonical correlation". The first canonical correlation is at least as high as the multiple correlation between the $q$ classes and one of the $p$ variables, and can be high even if all multiple correlations are small. In other words the linear combination defined by the coefficients $c_1, c_2, ..., c_p$ can show significant differences between the classes even if none of the variables within the model does.

## 2.2 Classification Trees (CT)

During the learning phase of the CT approach the set $\mathcal{E}$ of events $\vec{e}_k, (k = 1, \ ..., \ n)$ is split by repetitive cuts on a single variable resulting in successive subsets $\mathcal{E}_1, \ \mathcal{E}_2, \ \cdots$ beginning with $\mathcal{E}$ itself (fig. 1). These sequential cuts lead to an architecture called classification tree. Such a tree provides a hierarchical type of representation of the data space which can be used as base for the classifier by following the appropriate branches, i.e., by applying the successive cuts to an arbitrary event. The subsets of events $\mathcal{E}_2$ and $\mathcal{E}_3$ are disjoint, with $\mathcal{E} = \mathcal{E}_2 \bigcup \mathcal{E}_3$, similarly $\mathcal{E}_4$ and $\mathcal{E}_5$ are disjoint with $\mathcal{E}_2 = \mathcal{E}_4 \bigcup \mathcal{E}_5$, and $\mathcal{E}_3 = \mathcal{E}_6 \bigcup \mathcal{E}_7$. The subsets which are not split are called terminal subsets (rectangular boxes). These terminal subsets provide a partition of $\mathcal{E}$, a class label is associated to each terminal subset. There may be two or more terminal subsets with the same class label.

To explain how the split is made at each node, let us consider two classes $C_i$ and $C_j$. If $f_{i,j}(t)$ are the two associated continuous density functions, the distribution

$$F_{i,j}(x) = \int_0^x f_{i,j}(t)dt$$

is used to define the Kolmorov-Smirnoff distance

$$D(x) = |F_i(x) - F_j(x)|.$$

Let us consider an arbitrary variable $x_{l,(l=1,\cdots,p)}$. Let $x_l^*$ be the value of $x_l$ which minimizes the Kolmogorov-Smirnoff distance. Two subsets, purer in one class than the parent set, can be obtained by comparing $x_l$ to $x_l^*$ for each event. This minimizes the mean cost of misclassification according to Baye's rule [10].

For the $p$ variables associated with each event, $p$ Kolmogorov-Smirnoff distances are computed at each node, and we take

$$D(x_l^*) = Max_{xl}|F_i(x_l) - F_j(x_l)|.$$

The classification of real events is then straightforward. One event belonging to an unknown class is processed through the tree and its classification depends on the label of the terminal node it ends up.

## 2.3 Neural Networks (NN)

In the methods described above the derived classifier was a linear function of the variables. Such techniques fail in separating classes which are not linearly separable. A way to solve non-linear problems is the use of neural network methods. Neural networks are data processing architectures constructed from a large number of highly interconnected formal neuron units (fig. 2). Two of them, the "multilayered perceptron" and the "learning vector quantization", are described subsequently.

### 2.3.1 General Description of a Multi Layered Perceptron (MLP)

The first modelization of a formal neuron has been proposed by McCulloch and Pitts in [11]. Using the basic features of a biological neuron, they proposed a modelization in which each unit computes its output by performing a sum over all its input features, weighted by some coefficients $W_{ij}$, called the "synaptic strength". The corresponding output $y_i$ of an arbitrary neuron is obtained by a state transition function acting on the weighted sum of its inputs $x_j$. Examples for transitions function are:

step function: $f(x) = -1$ for $x < x_0, f(x) = 1$ for $x > x_0$
sigmoïd function: $f(x) = a\ (e^{kx} - 1)/(e^{kx} + 1)$
stochastic function: $f(x) = 1/(1 + e^{-x/T})$.

Then:

$$y_i = f(\sum_j W_{ij} x_j)$$

The most commonly used neural networks architecture for solving non-linear problems is the multi layered feed forward networks trained by back-propagation of the errors. Usually a sigmoïd function is used to transform the inputs into the output. One or several hidden layers provide non-linearity. The network output is obtained by applying to the input layer the pattern (event) vector $\vec{e_k}$. The outputs of every neuron layer are propagated forward through the network and the output of the network is provided by the last layer neuron output.

The weights $W_{ij}$ are first determined by supervised learning, in which well classified patterns are presented in turn to the network. The network output $y_k$ is then compared to the expected one $d_k$ for each pattern $\vec{e_k}$. An error function

$$E_k = (d_k - y_k)^2$$

is minimized by updating the weights at each presentation. This is done by using the "gradient descent method" with an error back-propagation algorithm [12]:

$$W_{ij}^{new} = W_{ij}^{old} + \Delta W_{ij}$$

$$\Delta W_{ij} = -\eta \frac{\partial E_k}{\partial W_{ij}} \;+\; \alpha \Delta W_{ij}$$

When a limited sample of training patterns is available the set is repeatedly presented to the network. When an unlimited sample is available the training procedure will use a new pattern at each training step.

In the validation step the network performance, now with steady weights, is checked with a different sample of patterns.

## 2.4 Learning Vector Quantization (LVQ)

A network using the LVQ algorithm [13] is a nearest-neighbour classifier. In a MLP neural network there is only one output for a class. When the classes overlap in parameter space the separation of the events is complicated. The aim of the LVQ neural network is to improve the separation by increasing the number of outputs related to each class.

An output layer neuron $i$ computes the distance between an arbitrary input $\vec{e}_k$ and a weight vector $\vec{W}_i = (W_{i1}, W_{i2}, ..., W_{i,p})$ (fig. 3). The output $S_i$ of this neuron $i$ is:

$$S_i = ||\vec{e}_k - \vec{W}_i||^2.$$

We call $C$ the nearest neuron of $\vec{e}_k$. During the learning phase the modification of the weights $\vec{W}_c$ is done according to:

- if class $(\vec{W}_c)$ = class $(\vec{e}_k)$

$$\vec{W}_c\,(t+1) = \vec{W}_c(t) + \alpha(t)\,(\vec{e}_k - \vec{W}_c(t))$$

- if class $(\vec{W}_c) \neq$ class $(\vec{e}_k)$

$$\vec{W}_c(t+1) = \vec{W}_c(t) - \alpha(t)\,(\vec{e}_k - \vec{W}_c(t))$$

- for the other vectors

$$\vec{W}_c(t+1) = \vec{W}_c(t)$$

where $\alpha(t)$ is a tuning parameter.

The network output gives the closest vector $\vec{W}_c$ of each validation event, and thus the class this event belongs to.

# 3 Selection of Input Variables

Although there is no restriction on the number of variables that are used for the classifier it is obvious that a small number will lead to a more manageable and less time consuming learning. Various methods exist to qualify the usefulness of a single variable with respect to its discriminating power and its correlation to other variables. We have used two methods to find appropriate subsets of variables as a basis for the multivariate methods:

- the $F$-test
- the stepwise selection.

## 3.1 The $F$-test

We consider a set $\mathcal{E}$ of $n$ events $\vec{e}_i$ divided into $q$ classes, and described by $p$ variables.

From the $e_{ij}$ values $i = 1,...,n,\ j = 1,...,p$ we define for an arbitrary variable $j$ $g_T(j)$ the barycenter of the whole event sample, and $g_\ell(j)$ the barycenter of events belonging to the arbitrary class $C_\ell$ with $n_\ell$ events

$$\begin{aligned} g_T(j) &= \frac{1}{n}\sum_{i=1}^{n} e_{ij} \qquad && j = 1,...,p \\ g_\ell(j) &= \frac{1}{n_\ell}\sum_{i\in C_\ell} e_{ij} \qquad && \ell = 1,...,q. \end{aligned}$$

It is useful to introduce the "within" vector $W$ describing the dispersion within a class

$$W(j) = \sum_{\ell=1}^{q}\sum_{i\in C_\ell}\frac{1}{n}\left(e_{i\ell} - g_\ell(j)\right)^2$$

and the "between" vector $B$ describing the distance of a class to the overall barycenter $g_T(j)$

$$B(j) = \frac{1}{n}\sum_{\ell=1}^{q} n_\ell \left(g_\ell(j) - g_T(j)\right)^2 .$$

Large values of $B(j)$ and small values of $W(j)$ characterize well separated and compact classes. Therefore the discriminating power of a variable $j$ is summarized in the $F$-test [14]

$$F(j) = \frac{n-1-q}{q-1}\,\frac{B(j)}{W(j)}$$

We start with the variable having the highest $F$-test value. Before adding a new variable we check its correlation with the variables previously selected and we ignore highly correlated variables.

### 3.2 Stepwise Selection (StS)

StS [15] is a method to reduce the dimensionality of multivariate analyses, which takes into account both the discriminating power of each single variable and the correlations between the variables. As a basis for STS we have used Wilks' $\Lambda$–statistic [16].

Defining the between and within class vectors as described above, Wilks' $\Lambda$–statistic reads

$$\Lambda(j) = \frac{|W(j)|}{|W(j) + B(j)|} \qquad j = 1, 2, ..., p.$$

In order to study the influence of one more variable it is useful to define the so called "partial $\Lambda$–statistic"

$$\Lambda(j+1) = \frac{\Lambda(1, 2, ..., p, p+1)}{\Lambda(j)}.$$

It has been shown [17], that the corresponding $F$–statistic is given by

$$F = \frac{n-q-p}{q-1} \cdot \frac{1-\Lambda(j+1)}{\Lambda(j+1)} \ .$$

That $F$–statistic is used to test the significance of the change from $\Lambda(j)$ to $\Lambda(j+1)$, which is a test of the improvement in the discriminating power by introducing another variable.

At each step of the StS procedure it is examined whether the variable in the model which contributes least to the discriminating power meets the criterion to stay. If the variable fails then it is removed. Otherwise the variable not yet in the model that contributes most to the discriminating power is entered. Both the criterion to stay and the criterion to enter are based on significance levels of the $F$–test deduced from the above defined $F$–statistic. By varying the significance levels one can control the number of variables in the model.

StS begins with no variable in the model. Therefore the variable with the highest discrimination power is entered first. The process stops when all variables in the model meet the criterion to stay and none outside the model meets the criterion to enter. The variables remaining in the model build the base of the multivariate analyses.

## 4 Tagging of $b$-Quark Events at LEP/SLC using Multivariate Analyses

Starting from a large set of variables which characterize the different event topologies of $b\bar{b}$ and light quark events ($c\bar{c}$, $s\bar{s}$, $d\bar{d}$, $u\bar{u}$) the various methods define a subset of variables by using the above described selection methods. Afterwards a classifier is derived by using multivariate analyses. This classifier is used to tag $b$-quark events.

### 4.1 Event Simulation

For the learning and validation step we have used about 600,000 $e^+e^- \to Z \to q\bar{q}$ Monte Carlo events produced with the ALEPH Monte Carlo program HVFL. The simulation is based on DYMU [18] and JETSET 7.3 [19] including the fine tuning of the parton shower parameters [20] and a special decay package which takes into account the current knowledge on beauty and charm physics. Detector effects have been introduced by using the standard ALEPH detector simulation program [21].

The four-momenta of the particles seen inside the detector are reconstructed by an "energy flow" algorithm, developed within ALEPH [22]. The algorithm combines information from the track chambers and the calorimeters. The reconstructed four-momenta are used to derive the various variables and to find the jets in the event using the JADE scaled-invariant-mass clustering algorithm [23] The parameter $Y_{cut}$ is set to $(M_{jet}/E_{vis})^2$. $E_{vis}$ is the total reconstructed energy. For $M_{jet}$, 6 $GeV/c^2$ was chosen, to gain the best jet axis resolution [24].

## 4.2 Physics variables

The larger mass of the $b$-quark with respect to other quark flavours has three major consequences: $b$-quarks loose less energy by gluon bremsstrahlung than light quarks, their fragmentation is harder and their decay products are more energetic. Thus the fraction of the beam energy carried by $B$-hadrons is 70% on average and only 51% for $D$-hadrons produced in $c\bar{c}$ events [25], resulting in different topologies for $b\bar{b}$ and light quark events. In particular $b\bar{b}$ events will appear more spherical than light quark events and the particles produced in $b\bar{b}$ events will have on average higher momenta and transverse momenta with respect to their jet axis.

Taking advantage of these characteristics we have defined a set of 70 purely kinematic variables. Two different types have been used: variables based on the full event shape, like sphericity and aplanarity, and variables based on the properties of the jets in the event, like the invariant mass of the most energetic jet. Details on the definition of the different variables have been given elsewhere [5].

## 4.3 Learning Step and Validation

Because the procedure is similar for all methods we will give details for MLP only.

The structure of the MLP used for this analysis is the following: one input layer with 9 neurons, two hidden layers with 9 and 6 neurons, respectively, and one final layer with 1 output neuron. The parameter $\alpha$ is set to 0.5, while $\eta$ can vary between 0.001 and 0.03 during the learning phase. The number of hidden layers, the number of neurons per layer and the values for $\alpha$ and $\eta$ have been optimized with respect to the best gainable separation between $b$- and non-$b$-events. The selection of the nine variables $A(I)_{F-\text{value}}$ $(I = 1, ..., 9)$, listed subsequently, used as inputs for the first layer was done with the $F$-test method described above.

- $A(1)_{989}$ is the boosted hemisphere sphericity product with $\beta_{boost} = 0.96$ [*].
- $A(2)_{898}$ is the sum of the products of the transverse momenta and the longitudinal momenta with respect to the jet axis normalized to $P^2_{total}$, where $P_{total}$ is the sum of the momentum of all the tracks in the event.

[*] This value optimizes the $b$-separation [26].

- Defining the total transverse jet momentum as the sum of the momentum components of the particles in the jet perpendicular to the jet axis, then $A(3)_{733}$ is the sum of the total transverse momenta of the jets of the event normalized to $P_{total}$.
- $A(4)_{343}$ is the momentum of the leading particle of the event normalized to $P_{total}$.
- $A(5)_{298}$ is the invariant mass of the three most energetic particles of the most energetic jet.
- $A(6)_{265}$, $A(7)_{166}$, $A(8)_{249}$, $A(9)_{219}$ are directed sphericities [4].

The MLP learning step has been performed with 9000 $b\bar{b}$, 9000 $c\bar{c}$ and 9000 $u\bar{u} + d\bar{d} + s\bar{s}$ fully simulated Monte Carlo events. The initial weights are chosen randomly between [-0.01,0.01]. One event of each class is processed through the network. We feed forward and back-propagate the error. The reactualisation of the weights is done after one exposure of an event of each class by minimizing the cost function $E$.

The learning procedure is stopped when the performance of the network ceases to improve significantly, i.e., when the function $E$ reaches asymptotically a minimal value. This corresponds roughly to 2 millions exposures (about 60 minutes of CPU time on IBM 3090). We have checked that a change of the relative fractions of $b\bar{b}$ from 33% to 50% in the learning event sample and the order in which we present the different event classes does not bias the result of the learning.

The discrimination power of the various methods between $b$-quark and light quark events was studied with our sample of 600.000 Monte Carlo events, excluding those events that have been used in the learning step. For each $q\bar{q}$ event the classifier output has been computed. The shape of the corresponding outputs are shown in fig. 4 for CDA and MLP. The different shape for $b$- and light quark events illustrates the ability of the multivariate approach to discriminate the different event classes effectively. Applying cuts on the classifier output provides $b$-enriched event samples. Fig. 5 compares the $b$-purity of the remaining sample as function of the efficiency of the applied cut, that can be reached with the various multivariate methods, together with the one reachable with the hemisphere boosted sphericity product only. MLP gives the best overall discrimination.

## 4.4 Application to the Measurement of $\Gamma(Z \to b\bar{b})$

As an example for an experimental application of multivariate analyses, we discuss in this section the measurement of the partial width of the $Z$ into $b\bar{b}$ at LEP/SLC.

### 4.4.1 The Single Tag Method

The shapes of the classifier output for $b$- and light quark events are parametrized by using a large sample of fully simulated events. These functions $f_b$ and $f_{light}$ are then normalized and the classifier output of the data $f_{data}$ is fitted according to the formula

$$f_{data} = N_b f_b + (N_Z - N_b) f_{light}$$

where $N_Z$ is the number of hadronic events in our sample and $N_b$ is the free parameter of the fit. From the fitted value we get

$$R_b = \frac{\Gamma(Z \to b\bar{b})}{\Gamma(Z \to hadrons)} = \frac{N_b}{N_Z}.$$

Such an analysis was already presented by the ALEPH collaboration [5] by using CDA and MLP.

The gainable statistical error of $R_b$ obtained from 400,000 $Z \rightarrow q\bar{q}$ events is very small

$$\Delta R_b / R_b \simeq 1\%$$

if we neglect the statistical uncertainty due to the limited number of simulated events. Comparing this value with the "theoretical" error of 0.43%, that would be obtained with an ideal separation between $b$- and light quark events, one proves the high discrimination power of the developed multivariate analyses.

However since the shape of the discriminator output for $b$- and light quark events is parametrized with simulated events this method is model sensitive. For instance the simulation of hadronization depends on phenomenological models which give rise to systematic uncertainties.

Several checks have been done to test the validity of the methods. In particular we have verified that the fine tuning of QCD Monte Carlo parameters is not sensitive to large variations of $R_b$, and therefore does not bias our analysis [25, 27]. Furthermore the shape of the classifier output for $b\bar{b}$ events has been checked by using real $Z \rightarrow q\bar{q}$ events containing leptons.

Typically we obtain an overall relative systematic error of about 5% with this single tag method, which is a serious limitation.

### 4.4.2 The Double Tag Method

We have developed another method which is less Monte Carlo dependent. This study has been done with MLP. The events are split into two hemispheres according to the plane perpendicular to the thrust axis and the network is "fed" with observables A(I) from each hemisphere. The structure of the network used for this study is the same as defined previously, but the 9 variables are computed for each hemisphere separately.

For a given cut on the MLP output, we define 3 classes of events: those where only one hemisphere is used as a tag ("single tagged events"), those where both hemispheres are required to satisfy the cut ("double tagged events") and events tagged by a high $p_\perp$ lepton on one side and by a cut on the MLP output on the other side ("single tagged high $p_\perp$ leptons"). This allows to extract from data $R_b$ and the cut efficiencies $\epsilon_{b\bar{b}}$ and $\epsilon_{light}$ for $b\bar{b}$ and light quark events $udsc$, provided the $b\bar{b}$ purity $P_{b\bar{b}}$ in the high $p_\perp$ lepton sample is given by the simulation [†], by solving the following system of equations:

$$\left\{ \begin{array}{lll} \epsilon^{ST}_{q\bar{q}} & = R_b \epsilon_{b\bar{b}} & + (1 - R_b)\epsilon_{light} \\ \epsilon^{DT}_{q\bar{q}} & = R_b \epsilon^2_{b\bar{b}}(1 + C^{DT}_{b\bar{b}}) & + (1 - R_b)\epsilon^2_{light}(1 + C^{DT}_{light}) \\ \epsilon^{ST}_{lepton} & = P_{b\bar{b}} \epsilon_{b\bar{b}} & + (1 - P_{b\bar{b}})\epsilon_{light}(1 + C^{ST}_{light}) \end{array} \right. \qquad (1)$$

$\epsilon^{ST}_{q\bar{q}}$ , $\epsilon^{DT}_{q\bar{q}}$ and $\epsilon^{ST}_{lepton}$ are the selection efficiencies for the single tagged events, the double tagged events and the single tagged high $p_\perp$ leptons, respectively.

$C^{ST}_{light}$ is a correction coefficient introduced to account for the fact that the $c\bar{c}$ events are worse separated from $b\bar{b}$ events than the $uds$-events (fig. 6). Since the $c\bar{c}$ fraction

[†] Note that $P_{b\bar{b}}$ can be in principle extracted from the data by using a global analysis of events with prompt leptons.

in the light quark sample $udsc$ is not the same for single tagged events and for single tagged high $p_\perp$ leptons, the cut efficiencies for $udsc$-events will be different in the two samples. In fact in the single tagged events sample we have 17% $c\bar{c}$ events and 61% $uds$ events while in the single tagged high $p_\perp$ leptons sample we have 7.7% $c\bar{c}$ events and 5.1% $uds$ events for a $p_\perp$ cut at 0.8 GeV.

$C^{DT}_{b\bar{b}}$ and $C^{DT}_{light}$ are correction factors which take into account possible correlations between the two hemispheres.

The coefficients $C^{ST}_{light}$, $C^{DT}_{b\bar{b}}$, $C^{DT}_{light}$ and the $b\bar{b}$ purity of the leptonic sample have to be estimated by Monte Carlo. We obtain $C^{ST}_{light} = (14. \pm 4.)\%$, $C^{DT}_{b\bar{b}} = (0.15 \pm 0.28)\%$, $C^{DT}_{light} = (2.95 \pm 0.62)\%$ and $P_{b\bar{b}} = (87.2 \pm 0.4)\%$, where the errors are due to our limited Monte Carlo statistics. Systematic errors on $C^{ST}_{light}$ have been studied by varying the normalization and the shape of the classifier output of the $c\bar{c}$ contribution. A variation of ±15% of the $c\bar{c}$ partial width gives an error of ±0.4% on $C^{ST}_{light}$. Varying for $c\bar{c}$ the $c$ quark fragmentation parameter $\epsilon_c$ from 0.020 to 0.060 [28] results in a ±1.2% error on $C^{ST}_{light}$. Adding all these errors in quadrature, we finally obtain $C^{ST}_{light} = (14. \pm 4.(stat.) \pm 1.3(syst.))\%$. A similar analysis gives a typical systematic error on $P_{b\bar{b}}$ of the order of 1%.

To illustrate the usefulness of the method, we have applied the analysis to a sample of 400,000 simulated $Z \to q\bar{q}$ events treated as if they were real data[‡]. With a cut at 0.4 on the MLP output and 0.8 GeV/c on the $p_\perp$ of the lepton, we obtain:

$$R_b = 0.212 \pm 0.008_{stat.} \pm 0.005_{syst.} \ .$$

The fraction of $b\bar{b}$ events in the sample was 0.22.

The systematic error on $R_b$ stems from the uncertainties in $P_{b\bar{b}}$ and in the correlation coefficients previously discussed. The different contributions are given in tab. 1.

The double tag method reduces the systematic error on $R_b$ by a factor 3 compared to the single tag method, at the expense of a worse statistical error due to the limited number of high $p_\perp$ leptons. Note that the understanding of all the possible sources of correlations is not an easy task and will rely on the simulation.

However this double tag method appears to be the best one in view of a high precision measurement of $R_b$ since it is less systematically limited than the single tag method. Furthermore the double tag method can be applied to any problem in which one needs to select a sample of events without introducing a priori any bias in the analysis of this sample. For instance the sample of events can be selected by applying a MLP cut on one hemisphere and then can be analysed by using the opposite hemisphere. This principle has been applied by the ALEPH collaboration to the study of $D^*$ production in $Z \to c\bar{c}$ [28].

# 5 Conclusion

Analysis of the shape of $Z \to q\bar{q}$ events allows to extract the $Z \to b\bar{b}$ fraction by using the information originally contained in the quark mass. This was first done by using

[‡] This statistic corresponds roughly to the number of $Z$ events collected by each LEP experiment at the end of 1991.

only the product of boosted sphericities [29] . The $b$-quark mass information is diluted in the event and the flavour tagging can be improved by using a larger set of sensitive variables. This needs more sophisticated methods to analyse the events, all of them relying on multivariate analysis. A large spectrum of such methods has been discussed in this paper: Fishers' linear and canonical discriminant analysis, classification trees and 2 categories of neural networks. The improvement provided by the multivariate approach with respect to the single-variable analysis is very significant. While the performances are very similar to the one of lepton tagging at low efficiency, the methods allow to tag 50% of the $b\bar{b}$ events with a signal to noise ratio of 1. The comparison of the different approaches indicates that non-linear methods map in a better way the complexity of the events. The neural network techniques seem to provide the asymptotic limit for $b$ event-tagging at LEP, when only the quark mass information is used.

The tagging techniques have been used to study the $\Gamma(Z \to b\bar{b})/\Gamma(Z \to q\bar{q})$ measurement, which is an ambitious and complex problem related to electroweak $b$-physics at LEP/SLC. They allow to reduce the present statistical limitations of the analysis based on semileptonic decays of $B$-hadrons [26]. Results have already been presented by the ALEPH collaboration [5] by applying these methods on the global shape of events. The conceptual problem of the global method is the use of Monte Carlo simulation to get the shape of $b\bar{b}$ and non-$b\bar{b}$ events. This limits the quality of the result due to large systematic errors arising from modelization, which were evaluated to be about 5%. Unfortunately the large systematic error makes the result insensitive to electroweak effects which require a relative error of the order of 1%.

For this reason we have developed a different method for which the Monte Carlo simulation is mainly used for the learning step of the neural network; the tagging efficiencies for $b\bar{b}$ and light quark events are extracted directly from data by double hemisphere tagging. A systematic error at the level of 2.5 % is conservatively obtained. This kind of method could be made more efficient in the near future by adding the $b$ lifetime information in the multivariate analyses. Then the systematics on $\Gamma(Z \to b\bar{b})$ measurement could be decreased below 1%. This would be done at the expense of a smaller angular acceptance. This limitation is crucial for the $b$- and $c$-asymmetry measurements for which the $b$-tagging approach presented in this paper could be more performant. More generally this method can be used to tag events in one hemisphere and analyse physics in the opposite hemisphere. This has already been done by the ALEPH collaboration to study charm physics.

### Acknowledgements

We would like to thank our colleagues of the ALEPH collaboration for their support. Thanks are also due to Jacques Lefrancois for stimulating the double tag approach, to M$^{me}$ F. Fougelman-Soulié for useful discussions on neural networks and to G. Rudolf for his work on Monte Carlo fine tuning.

## References

[1] S.L.Glashow, Nucl. Phys. **22** (1961) 579.

S.Weinberg, Phys. Rev. Lett. **19** (1967) 1264.
A.Salam, Proc. $8^{th}$ Nobel Symp., editor: N. Svartholm, Stockholm (1968) 367.

[2] I.Guyon, Phys. Rep. **207** (1991) 215.
F.Fogelman-Soulié, Proc. Workshop on Neural Networks: From Biology to High Energy Physics, editor: D. Benhar *et al.*, Ets. Editrice, Pisa (1992).

[3] R.Odorico, Phys. Lett. **B120** (1983) 219.
G.Ballochi, R.Odorico, Nucl. Phys. **B229** (1983) 1.
M.Mjaed, J.Proriol, Phys. Lett. **B127** (1989) 560.
L.Lonnblad, C.Peterson, T.Rognvaldsson, LUTP-90-8.
C.Bortolotto *et al.*, Nucl. Inst. and Meth. **A306** (1991) 459.
C.Bortolotto *et al.*, Proc. Workshop on Neural Networks: From Biology to High Energy Physics, editor: D.Benhar *et al.*, Ets. Editrice Pisa (1992).
J.Proriol *et al.*, same reference as above.
B.Brandl *et al.*, HD-IHEP 92-1, "Tagging of $Z$ Decays into $b$–Quarks in the ALEPH Detector using Multivariate Methods: Discriminant Analyses, Artificial Neural Network", invited talk at the Second International Workshop on Software Engineering, Artificial Intelligence and Expert Systems for High Energy and Nuclear Physics, L'Agelonde, January 1992, to be published in the proceedings.
J.Jousset *et al.*, PCCF RI 9206, "Jets Recognition and Tagging with Neural Networks", same reference as above.
P. Branchini, M. Ciuchini, P. Del Guidice "B tagging with Neural Networks: an alternative use of single particle information for discriminating jet events" , same reference as above.

[4] L.Bellantoni *et al.*, Nucl. Inst. and Meth. **A310** (1991) 618.

[5] P.Henrard, ALEPH coll., Proc. $4^{th}$ International Symposium on Heavy Flavour Physics, editors: M.Davier, G.Wormser, Editions Frontières (1992).
B.Brandl, ALEPH coll., HD-IHEP 92-4, "Measurement of the Partial Width of the $Z$ into $b\bar{b}$" , invited talk at the XXVIIth Recontres de Moriond: Electroweak Interactions and Unified Theories, Les Arcs, March 1992, to be published in the proceedings.

[6] R.A.Fisher, Annals of Eugenics **7** (1936) 179.

[7] G.Celeux, "Analyse Discriminante sur Variables Continues", INRIA, Collection Didactique n° 7, Paris (1990).

[8] Y.Fujikoshi, in "Multivariate Analysis", Vol. 6, editor: P.R.Krishnaiah, North Holland, Amsterdam (1985).

[9] T.W.Anderson, "An Introduction to Multivariate Statistical Analysis", Wiley, New York (1958).

[10] L.Breiman *et al.*, "Classification and Regression Trees", Wadsworth, California (1984).
G. Celeux, Y. Lechevallier, COMPSTAT 82 p. 161.

[11] W.S.McCulloch, W.Pitts, Bull. of Math. Biol. **5** (1943) 115.

[12] D.E.Rumelhart, J.L.McClelland, "Parallel Distributed Processing ", Vol. 1, MIT Press (1988).

[13] T.Kohonen, "Self Organization and Associative Memory", Springer, New York (1989).

[14] M.S.Srivastava, E.M.Carter, "An Introduction to Applied Multivariate Statistics", North-Holland, Amsterdam (1983).

[15] R.I.Jennrich, in "Statistical Methods for Digital C,mputer", editor: K.Enslein *et al.*, Wiley, New York (1977).

[16] S.S.Wilks, "Mathematical Statistics", Princeton Univ. Press (1962).

[17] C.R.Rao, "Linear Statistical Interference and Its Applications ", Wiley, New York (1965).

[18] J.E.Campagne, Ph.D. thesis, LPNHEP-89-02.
J.E. Campagne, R.Zitoun, Z. Phys. **C43** (1989) 469.

[19] T.Sjostrand, Comp. Phys. Commun. **39** (1986) 347.
T.Sjostrand, M.Bengtsson, Comp. Phys. Commun. **43** (1987) 367,
M.Bengtsson, T.Sjostrand, Phys. Lett. **B185** (1987) 435,
T.Sjostrand *et al.*, "The Lund Monte Carlo Programs", CERN Pr. Libr. (1989) and references therein.

[20] D.Buskulic *et al.*, ALEPH coll., CERN-PPE-92-62.

[21] D.Decamp *et al.*, ALEPH coll., Phys. Lett. **B244** (1990) 551.

[22] D.Decamp *et al.*, ALEPH coll., Phys. Lett. **B246** (1990) 306.

[23] W.Bartel *et al.*, JADE coll., Z. Phys. **C25** (1984) 231 and Z. Phys. **C33** (1986) 23.
S.Bethke *et al.*, JADE coll., Phys. Lett. **B213** (1988) 235.

[24] D.Decamp *et al.*, ALEPH coll., Phys. Lett. **B263** (1991) 325.

[25] B.Brandl, Ph.D. thesis, HD-IHEP 91-7, and references therein.

[26] P.Roudeau, "Heavy Quark Physics at LEP", invited talk at the Lepton-Photon-Conference HEP, Geneva, August 1991, to be published in the proceedings.

[27] G.Rudolph, private communication.

[28] J.Boucrot, ALEPH coll., "Spectroscopy of $D$- and $B$-mesons in ALEPH", invited talk at the XXVIIth Recontres de Moriond: QCD and $B$-physics, Les Arcs, March 1992, to be published in the proceedings.
A. Bencheikh, PhD Thesis, Clermont-Ferrand (1992).

[29] P.Abreu *et al.*, DELPHI Coll., CERN-PPE-92-79

# Appendix

The following table gives the main software packages used to perform the multivariate analyses, the address of the authors and the used functions.

| Packages | Address | Functions |
|---|---|---|
| SAS | SAS Circle<br>CARY N.C. 27512-800 (USA) | Canonical discriminant analysis<br>General statistics<br>Stepwise selection |
| BMDP | University of California<br>Los Angeles CA (USA) | Selection of variables<br>Discriminant analysis |
| MODULAD | Club MODULAD<br>INRIA<br>Rocquencourt<br>F-78153 Le Chesnay Cedex | SELDISC: selection of variables<br>DISC: Fishers' linear discriminant analysis<br>MLP: multi layer perceptron<br>DNP: classification tree |
| JETNET 2 | Dr. Rogualdsson, Prof. C. Peterson<br>Department of Theoretical Physics<br>Lund University<br>S - LUND | General neural network packages<br>(preprint LUTP 91-18) |

| Source | Effect on $R_b$ |
|---|---|
| $P_{b\bar{b}}$ | $\pm0.0040$ |
| $C^{DT}_{b\bar{b}}$ | $+0.0016$<br>$-0.0015$ |
| $C^{DT}_{light}$ | $+0.0018$<br>$-0.0019$ |
| $C^{ST}_{light}$ | $+0.0017$<br>$-0.0014$ |

Table 1: Contributions to the systematic error on $R_b$.

## Figure Captions

Figure 1: Scheme of CT.

Figure 2: Scheme of MLP Neural Network.

Figure 3: Scheme of LVQ Neural Network.

Figure 4: CDA (a) and MLP (b) classifier ouputs for *b*- and light quark events.

Figure 5: The *b*-purity of the remaining sample as function of the efficiency of the applied cut for the various multivariate methods, together with the one reachable with hemisphere boosted sphericity product only.

Figure 6: Shape of the "one-hemisphere-MLP" classifier output for *c*- and *uds*-quark events.

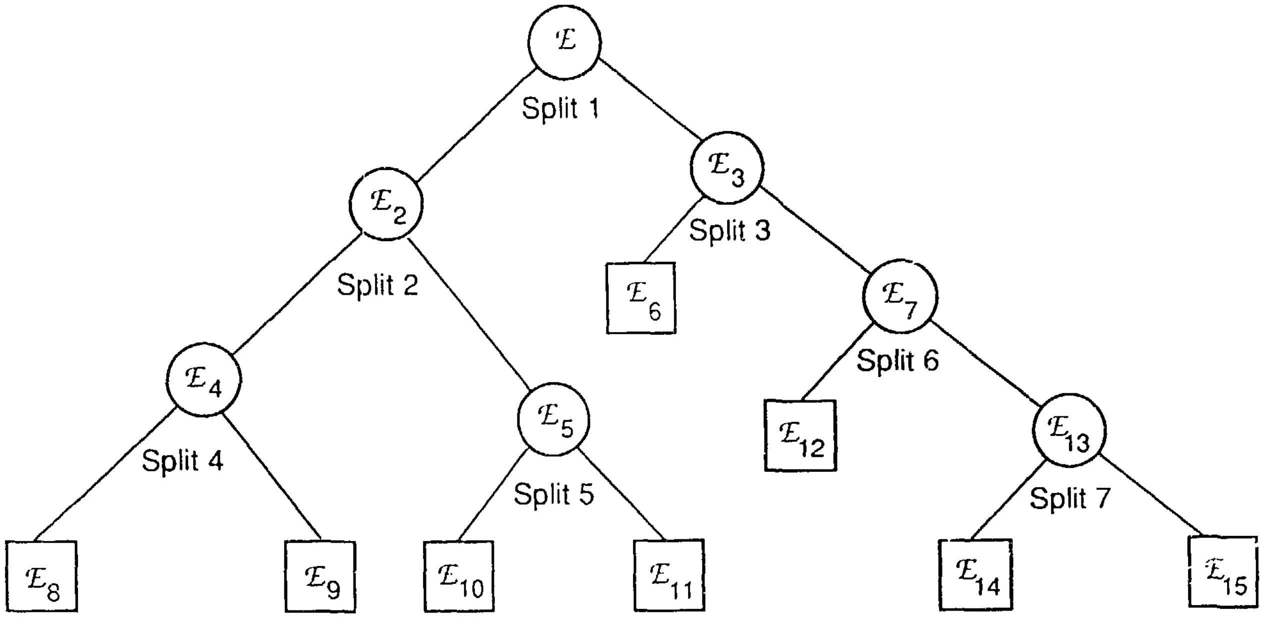
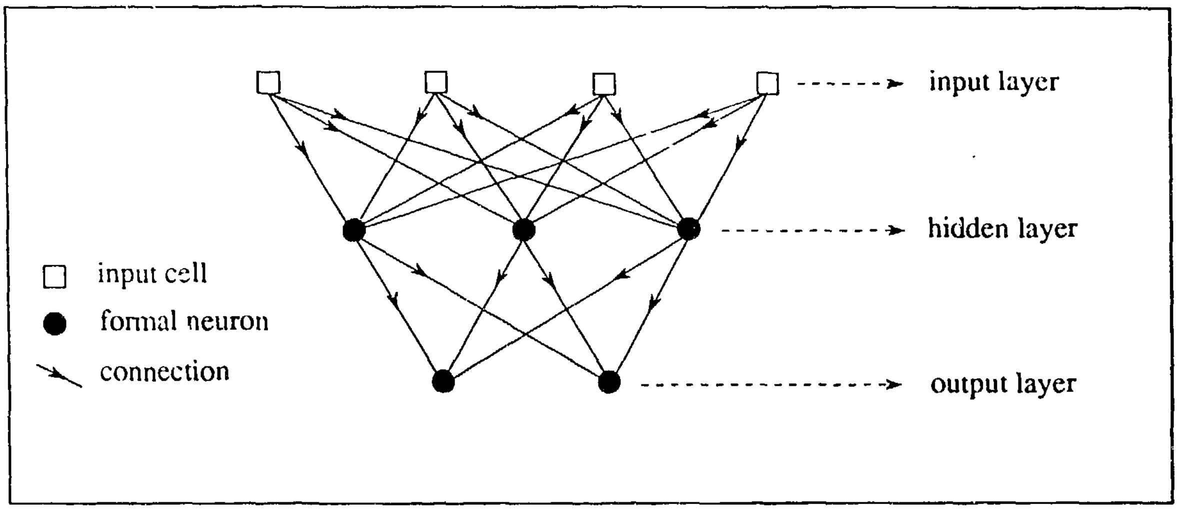
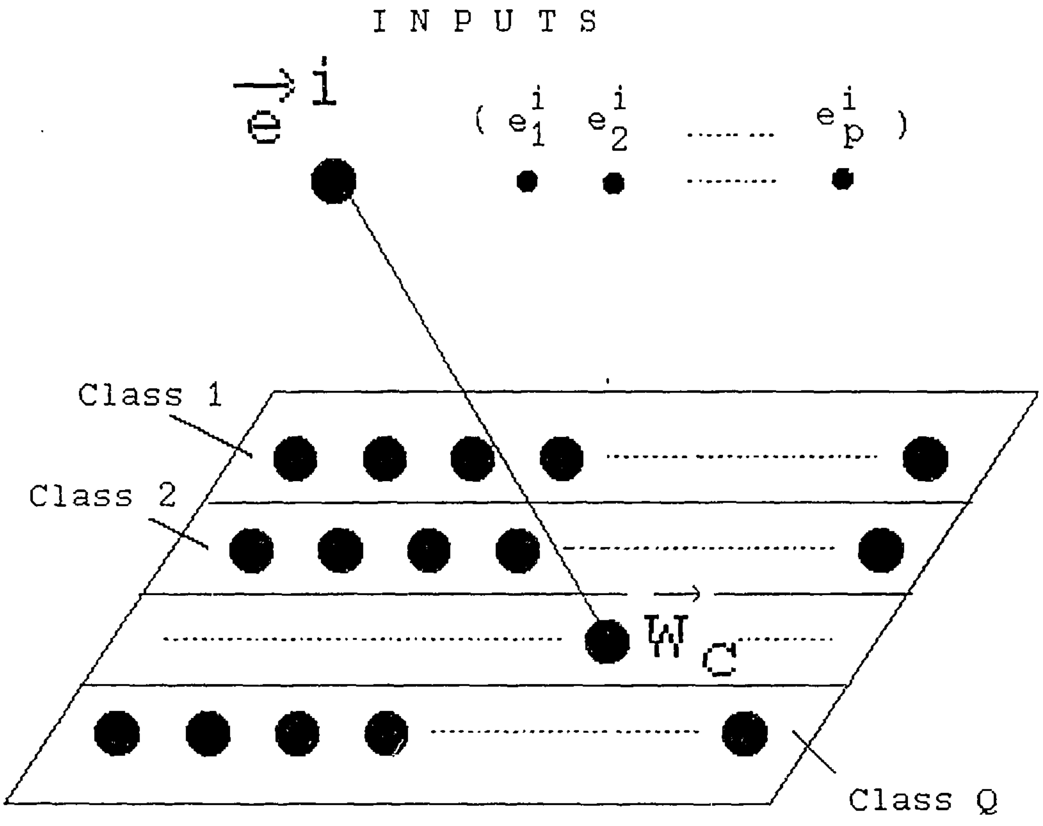

$\mathcal{E}$
Split 1
$\mathcal{E}_2$
$\mathcal{E}_3$
Split 2
Split 3
$\mathcal{E}_6$
$\mathcal{E}_7$
$\mathcal{E}_4$
Split 6
$\mathcal{E}_5$
$\mathcal{E}_{12}$
$\mathcal{E}_{13}$
Split 4
Split 5
Split 7
$\mathcal{E}_8$
$\mathcal{E}_9$
$\mathcal{E}_{10}$
$\mathcal{E}_{11}$
$\mathcal{E}_{14}$
$\mathcal{E}_{15}$

Fig. 1



Multilayered feed-forward network

Fig. 2



OUTPUT (Neurons mapping)

Figure 3

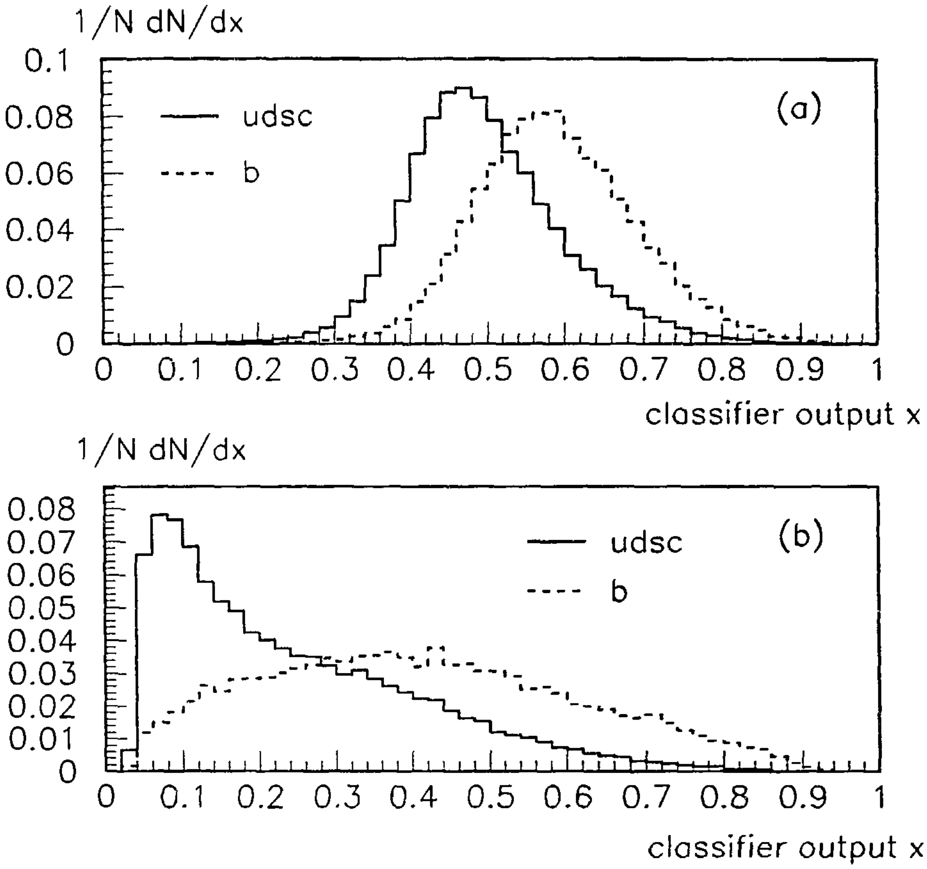

1/N dN/dx
0.1
0.08
0.06
0.04
0.02
0
udsc
b
(a)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
classifier output x
1/N dN/dx
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
udsc
b
(b)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
classifier output x

Figure 4

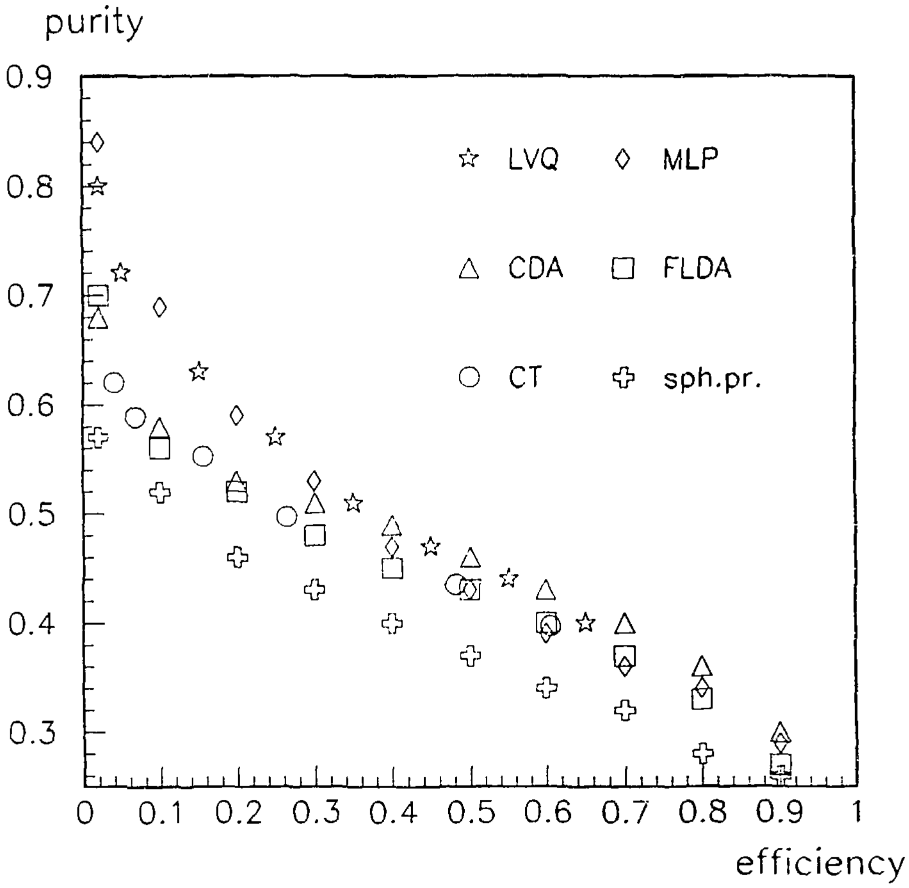

purity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
efficiency
☆ LVQ
◇ MLP
△ CDA
□ FLDA
○ CT
✣ sph.pr.

FIGURE 5

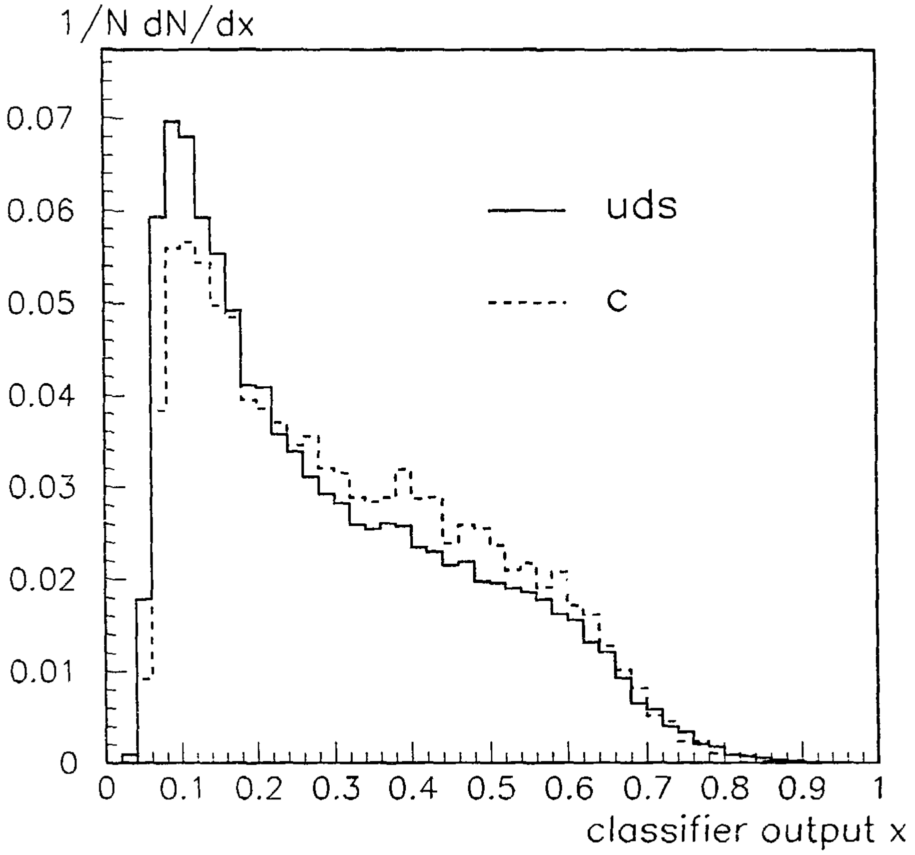

1/N dN/dx
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
uds
c
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
classifier output x

Figure 6