DEUTSCHES ELEKTRONEN-SYNCHROTRON

DESY 94-217
November 1994

Jet Identification Based on Probability Calculations Using Bayes' Theorem

C. Jacobsson, L. Jönsson, M. Nyberg-Werther

Physics Department, University of Lund, Sweden

G. Lindgren

Department of Mathematical Statistics, University of Lund, Sweden

ISSN 0418-9833

NOTKESTRASSE 85 - 22603 HAMBURG

DESY 94-217

November 1994

Jet Identification based on Probability Calculations using Bayes' Theorem

C. Jacobsson[a], L. Jŏnsson[a], G. Lindgren[b], M. Nyberg-Werther[a]

[a] Physics Department, Lund University, Sŏlvegatan 14, S-223 62 Lund, Sweden
[b] Department of Mathematical Statistics, Lund University, Box 118, 221 00 Lund, Sweden

Abstract

The problem of identifying jets at LEP and HERA has been studied. Identification using jet energies and fragmentation properties was treated separately in order to investigate the degree of quark-gluon separation that can be achieved by either of these approaches. In the case of the fragmentation-based identification, a neural network was used, and a test of the dependence on the jet production process and the fragmentation model was done. Instead of working with the separation variables directly, these have been used to calculate probabilities of having a specific type of jet, according to Bayes' theorem. This offers a direct interpretation of the performance of the jet identification and provides a simple means of combining the results of the energy- and fragmentation-based identifications.

ISSN 0418-9833

1 Introduction

In many tests of QCD, based on processes producing jets, it is of great importance to be able to identify whether a jet originates from a quark or a gluon. Different criteria such as specific decay properties, prior knowledge of the short range dynamics of the process or differences in the topology of jets due to hadronisation can be used in such an identification.

Especially heavy quarks can be identified from their decay properties by using apropriate particles in the decay chain to tag the flavour of the heavy quark. Fast leptons from semileptonic decays have been used, as well as charged kaons and D-mesons. Further, the long decay time of weak decays offers the possibility of reconstructing the secondary decay vertex by using high resolution vertex detectors.

The short range dynamics defines the kinematic properties of the process. For example, the fact that gluons are produced from primary quarks in a bremsstrahlung-like process implies that the gluon jets are usually less energetic than the quark jets in an event.

The topology of jets is due to features of the partons which are related to their intrinsic properties, such as mass and colour charge. Such differences influence the way the partons fragment into final state hadrons forming jets.

In this analysis we have studied jet separation from a general aspect and therefore have concentrated on differences in the jet energies and in those properties of jets that are related to the fragmentation process. The optimal cut in jet energy for a separation between quarks and gluons obviously depends on how much energy is available for a certain process and how many jets are produced in that process. In e^+e^- collisions the energy available for jet production is well defined, while for ep collisions the energy involved in the hard scattering subprocess varies from event to event. For the description of the shape of jets a large number of fragmentation variables are available. In principle the fragmentation of a parton should not depend on the way it has been produced, i.e. if the underlying process is an e^+e^- or an ep collision, especially if we restrict ourselves to consider the jet core which makes a possible influence of the colour strings less important. We have thus made an attempt to find a process-independent method to identify quarks and gluons by using suitable fragmentation variables alone.

The neural network method has previously been used with various input variables for the purpose of separating gluon jets from quark jets [1]. For various reasons which will be explained in the following, we have in this analysis distinguished between identification based on jet energies and identification based on fragmentation properties, using a neural network in the latter case. In a final step the two methods have been combined in order to improve the result by using all the available information. This can easily be done if, instead of working with the various separation variables directly, one converts these into probabilities that a jet is a quark or a gluon jet and applies cuts in the combined probabilities.

Most previous attempts to perform jet identification have been based on studies of individual jets. The approach which we have adopted is to isolate the specific event type of interest and use the additional information contained in the knowledge of the exact number of quarks (q), antiquarks ($\bar{q}$) and gluons (g) for that event. Experimentally this method is possible for 3-jet events in e^+e^- collisions which must be of the type $q\bar{q}g$, whereas it is a good approximation to assume that 4-jet events consist of a $q\bar{q}gg$ configuration, since the $q\bar{q}q\bar{q}$ contribution is strongly suppressed. In the case of ep collisions (2+1) jet events denote events with two jets in the hard scattering system in addition to the jet from the spectator quarks. The spectator jet, which is easily identified, is of no direct interest in the study of the hard subsystem which is the motivation for treating such events as 2-jet events in the following. The final state of these events is either of qg type (the QCD-Compton process) or $q\bar{q}$ type (the Boson-Gluon fusion process), although in certain kinematic regions the $q\bar{q}$ events can be neglected and we are left with a clean sample of qg events. For higher jet multiplicities the situation becomes less clear and an event-based identification can not be easily applied. In this study we have thus concentrated on 3-jet events generated at the LEP centre-of-mass energy, 91.2 GeV, and (2+1) jet events of the qg-type from simulated collisions between 820 GeV protons and 26.7 GeV electrons at HERA. The Feynman diagrams for the processes investigated here are shown in Fig. 1.

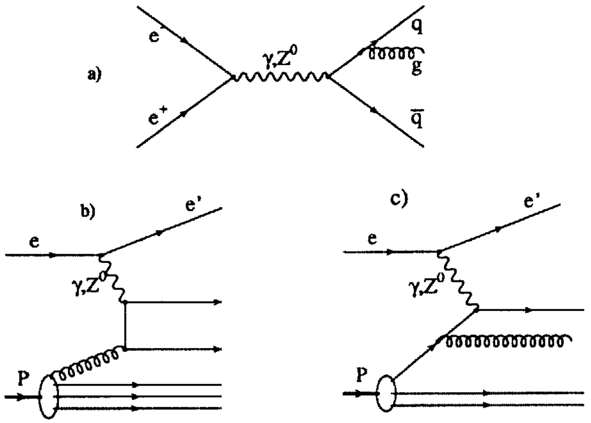

Figure 1: *The Feynman diagrams for a) a 3-jet event from an e^+e^- collision, b) a BGF event from an ep collision and c) a QCD-Compton event from an ep collision.*

2 Event generation

In order to investigate whether our jet identification based on fragmentation properties is process-independent, we have applied our method to both Monte Carlo-generated e^+e^- events and ep events. The generation of ep events has been done with the Monte Carlo (MC) programs LEPTO and HERWIG, which are known to reproduce Deep Inelastic Scattering (DIS) data from previous fixed target lepton-nucleon scattering experiments and also to give a fair description of the limited data on jet physics currently available at HERA. The Monte Carlo program JETSET has proven to give a good description of various e^+e^- data and was thus used to produce such events.

The basic concept of the event generators is that hard scattering processes can be factorized into an elementary hard process, initial- and final state radiation, and a hadronisation process. This general scheme can be used to describe a large variety of QCD and electroweak processes by applying different elementary subprocess matrix elements.

The JETSET program [2] describes e^+e^- annihilation into hadronic final states, using two alternative approaches. One is the calculation of explicit matrix elements (ME) up to the second order in α_s and the other is based on parton shower (PS) emission, which allows the production of an arbitrary number of jets. Although both methods have their advantages and disadvantages we have, in this analysis, chosen the parton shower option based on the coherent evolution scheme by Marchesini and Webber. A parton shower is based on the branchings $q \to qg, g \to gg$ and $g \to q\bar{q}$ as given by Altarelli-Parisi evolution equations in the leading logarithm approximation of perturbative QCD. The evolution is performed in an iterative manner and stopped when all parton masses have evolved below some minimum mass. This leads to an ordering in angle in the sense that angles between two emitted partons decrease with consecutive branching. The hadronisation is performed according to the Lund string model [3] [2].

LEPTO [4] simulates the basic DIS neutral and charged current processes and we have chosen to consider only the dominant neutral current process where a reconstruction of the event kinematics from the scattered electron can be made. In addition to the leading order Quark Parton Model (QPM) process, $\gamma^*q \to q$, where no jet identification is needed and therefore is of no interest for this analysis, also first order (α_s) processes, i.e. the QCD-Compton process, $\gamma^*q \to qg$, and the boson-gluon fusion process, $\gamma^*g \to q\bar{q}$, are calculated from QCD matrix elements. In order to avoid divergences from soft and collinear parton emission, a cut-off in the invariant mass of any two partons, m_{ij}, is implemented. Higher order corrections are then included by adding parton showers according to the same scheme, based on the Altarelli-Parisi evolution equation, as in JETSET. The amount of initial- and final-state radiation is determined by the virtual mass of the initiating parton just before and after the boson vertex. The initial-state radiation is generated by a backward evolution scheme from the hard vertex which is controlled by the parton density function specified in the program. We have used the MRS H parametrisation of the density function, as it describes recent results on the proton structure function F_2 at low Bjorken-x values from HERA. As in the case of the e^+e^- events, the Lund string fragmentation is used to produce the hadronic final state.

Similar to the LEPTO program, the simulation of (2+1) jet ep events by the HERWIG generator [5] is done using matrix elements to describe the first order processes in the strong coupling constant, and higher order emissions are introduced by parton showers which are generated essentially with Q^2 as mass scale. The upper limit for the shower evolution variables are related to energies and angles rather than to parton virtualities. The backward evolution process produces coherent, initial-state parton showers with full QCD cascading of all emitted partons. The same parton density function was used as for LEPTO. The final-state coherent showers include soft gluon interference and azimuthal correlations due to the gluon spin. The emitted gluons are split into quark antiquark pairs (or possibly into diquark antidiquark pairs) between which there are colour lines forming colour-singlet clusters. The clusters created in this way are fragmented into hadrons through a longitudinal splitting of the high mass clusters and phase space decay of lower mass clusters. This fragmentation scheme is normally denoted cluster fragmentation.

As already mentioned, this study has been limited to simulated 3-jet events from e^+e^--interactions at LEP and (2+1) jet events from generated ep collisions at HERA. No detector simulation has been done but, for the analysis of HERA events, the usual beam pipe cut was introduced, excluding the regions in polar angle below 4° and above 176° not covered by the detector. Due to the event topologies of deep inelastic scattering processes most of the spectator jet will disappear undetected down the forward cone, while for the majority of events the scattered electron will proceed inside the backward cone. No such cut is necessary at LEP since no specific activity is expected in these regions and since the jet analysis is in any case limited to the barrel region of the detector.

Since the Monte Carlo generators we have used cover jet production in e^+e^- as well as ep collisions and, in addition, use two different fragmentation schemes, we are able to test both the process- and model-dependence. This is done by training the neural network on samples from either of the generators and comparing the results when the network is applied to a test sample from the same generator and from one of the other generators, respectively, according to the following procedure. A comparison of the results from the network when trained on event samples from HERWIG and LEPTO respectively and subsequently tested on an event sample generated by LEPTO, will provide the model-dependence. On the other hand, if the respective event samples from JETSET and from LEPTO are used to train the network, which is then tested on a sample from JETSET, the process-dependence will come out. Finally, if the network trained on the samples from HERWIG and from JETSET respectively is applied on the test sample from JETSET, we obtain both the model- and the process-dependence.

3 Jet reconstruction

In order to reconstruct particle jets, the LUCLUS algorithm [6], based on the combination of energy clusters, was used. A careful study of the reconstruction quality as a function of the resolution parameter in the algorithm showed that a value d_{join} = 4 GeV was relevant (see [7]). In the HERA analysis the clustering was done in the so-called hadronic center-of-mass system i.e. in the center-of-mass system of the incoming proton and the exchanged virtual photon. In order to reconstruct the spectator jet in the best possible way, a pseudoparticle is added to each event to represent the fraction of the proton fragment lost in the beam pipe. The momentum of this pseudoparticle is given by the difference between the longitudinal momentum of the initial state and the measured longitudinal momentum of the final state, as described in [8].

In a Monte Carlo generator which includes parton showers, many partons can contribute to a jet and one needs a method to establish whether the reconstructed jet should be regarded as originating from a quark or from a gluon. For ep events, where the ME forms the basis of the processes and PS are added to simulate higher order corrections, we simply checked which reconstructed jet was closest to the original parton from the ME, according to: $\min(|\bar{P}_{jet1} - \bar{P}_q| + |\bar{P}_{jet2} - \bar{P}_g|, |\bar{P}_{jet2} - \bar{P}_q| + |\bar{P}_{jet1} - \bar{P}_g|)$. Since the e^+e^- events are not generated with ME, we have to extract the momentum vectors of the jets for both the parton level and the hadron level by applying the LUCLUS jet algorithm. A comparison of the momentum vectors

in pairs on the parton and hadron level, identified which jet on the hadron level corresponds to a certain jet on the parton level. If the jet on the parton level contains an odd number of quarks, it is considered to be a quark-initiated jet, while if the number of quarks in the jet is even, it is defined as a gluon jet. For almost all three jet events, the result of this definition is exactly one gluon and two quark jets. If not, the event is discarded.

4 Event selection

4.1 Selection of HERA events

Since we want to concentrate on the (2+1) jet events we have considered only those events in which the jet algorithm found exactly two jets + the spectator jet. In the hadronic centre-of-mass system we required the minimum energy of each jet to be 5 GeV and the invariant mass of the two jets to be larger than 15 GeV to ensure that the selected events had a reasonably clear jet structure. We also required a minimum of four particles to be assigned to each jet, since the jet variables which were used to study the fragmentation, are not meaningful for jets with too few particles. In order for the two hard jets to be well inside the acceptance region of the HERA experiments and to have a separation in space from the proton remnant, both jets had to be reconstructed within the region of polar angles $10° < \theta < 160°$ as measured in the laboratory system. The two jets also had to be separated by less than two units of pseudorapidity. It has been previously shown [7] that this is necessary in order to cut down the background of q-type events which otherwise enter the (2+1) jet sample.

Finally, we use only events produced within certain limits of the kinematic variables generally used to describe DIS events. These are Q^2, the momentum transfer squared, the Bjorken-x and -y scaling variables and W^2, the invariant mass squared of the hadronic system:

$$Q^2 \equiv -q^2 = -(p_e - p_\ell)^2, \quad x \equiv \frac{Q^2}{2P \cdot q}, \quad y \equiv \frac{P \cdot q}{P \cdot p_e}, \quad W^2 \equiv (q + P)^2 = Q^2\frac{1-x}{x} + m_p^2$$

(For a description of these variables and how they can be measured at HERA, see, for example, [9]).

The cross-section falls rapidly with increasing x and Q^2, for both the qg - and the $q\bar{q}$ type of events, but qg events dominate in the region of high-x-values. At $x \geq 0.1$ an almost pure sample of qg events is produced, $\frac{qg}{q\bar{q}} > 8$, and one can thus concentrate on separating quark jets from gluon jets in this region. A sample of qg events from this region was therefore used in our attempts to identify gluon jets at HERA. At lower values of x ($x < 0.1$) a mixture of qg - and $q\bar{q}$ events is produced and therefore the problem of separating the two event types has to be dealt with.

4.2 Selection of LEP events

In the selection of Monte Carlo-generated 3-jet events from e^+e^- collisions at LEP, we required each jet to have an energy of more than 5 GeV in order to have reasonably collimated flows of particles. Exactly as for the jets from ep collisions, the invariant mass of any jet pair, m_{ij}, should exceed 15 GeV to give an observable 3-jet topology. Also in analogy with the treatment of ep collision events we required each jet in an e^+e^- event to contain at least 4 particles, since the same fragmentation variables are going to be used in both cases. The energy sum of all three jets in an event was required to be greater than or equal to 90 GeV in order to guarantee that no fraction of a jet had escaped detection. We have assumed a LEP detector with full azimuthal coverage but restricted the jets to fall inside the range $40^\circ < \theta < 140^\circ$ of polar angle. This is the barrel region which is normally well covered by both the tracking system and the calorimetry of a detector.

5 Identification of gluon jets using jet-energy

Based on the assumption that gluon jets carry less energy than quark jets and using the known number of quarks and gluons in the event type under investigation, we want to calculate the probability that a jet originate from a gluon. In doing this we recall that there is a difference between e^+e^- collisions and ep collisions in the sense that the energy sum of quarks and gluons is constant and equal to $\sqrt{s}$ for e^+e^- processes, while the energy entering the hard scattering subprocess in ep collisions varies from event to event. An identification which includes jet energies will therefore always be process-dependent.

The probability of emitting a gluon with a certain energy in an e^+e^- collision is obtained directly from first order ME calculations. The conditional probability of jet1 in a 3-jet event to originate from a gluon, provided the jets have the energies E_1, E_2 and E_3, is given by:

$$P^{B_{123}}_{gqq}(jet1, jet2, jet3|E_1, E_2, E_3) = \frac{E_2^2 + E_3^2}{(E_{cm} - 2E_2)(E_{cm} - 2E_3)} \cdot \left(\frac{E_1^2 + E_2^2}{(E_{cm} - 2E_1)(E_{cm} - 2E_2)} + \frac{E_2^2 + E_3^2}{(E_{cm} - 2E_2)(E_{cm} - 2E_3)} + \frac{E_3^2 + E_1^2}{(E_{cm} - 2E_3)(E_{cm} - 2E_1)} \right)^{-1} \quad (1)$$

where E_{cm} is the centre-of-mass energy of the e^+e^- collision.

The probability that the scattered quark radiates a gluon of a certain energy in an ep-interaction is not so easily accessible from the matrix element and we have therefore used Monte Carlo-generated energy distributions to extract the density functions for the gluon and the quark, f_g and f_q, in a qg event.

$$f_g^B = \frac{f^B_{gq}(E_g, E_q)}{f^B_{g+q}(E_g + E_q)}$$

$$f_q^B = \frac{f^B_{gq}(E_q, E_g)}{f^B_{g+q}(E_g + E_q)}$$

where f^B_{gq} is the joint density function, given that the gluon and the quark have the energies E_g and E_q, respectively. The total energy of the quark gluon system, $E = E_g + E_q$, is a random variable with the density $f^B_{g+q}(E_g + E_q)$. Since we know that the *a priori* probability for selecting a quark or a gluon is equal, then the probability for jet1 to be a gluon jet is given by Bayes' theorem (see Appendix A):

$$P^{B_{12}}_{gq}(jet1, jet2|E_1, E_2) = \frac{0.5 f_g^B}{0.5 f_g^B + 0.5 f_q^B} = \frac{f^B_{gq}(E_1, E_2)}{f^B_{gq}(E_1, E_2) + f^B_{gq}(E_2, E_1)} \quad (2)$$

where E_1 and E_2 give the energies of jet1 and 2, respectively.

In each event the jet with the highest probability, $P_{g,max}$, is selected to originate from a gluon. The allowed range of gluon probabilities is, for 2-jet events $\frac{1}{2} < P_{g,max} < 1$, and for 3-jet events $\frac{1}{3} < P_{g,max} < 1$, where the lower bounds correspond to equal probabilities for all jets in the event to be a gluon. A general definition of the allowed range would thus be $\frac{1}{\#jets} < P_{g,max} < 1$, giving the limits within which a cut (P_{cut}) can be specified in order to enhance the purity of gluon jets in our selected sample. (For processes containing more than one gluon, the procedure can in principle be repeated to find a second gluon in the remaining sample and so on. The probability bounds for the second gluon will then be $\frac{1}{\#jets-1} < P_{g,max} < 1$ and for the i:th gluon candidate $\frac{1}{\#jets+1-i} < P_{g,max} < 1$.) From the generated MC data we can now check whether the jet with the value $P_{g,max}$ and thereby identified as a gluon jet, was actually initiated by a gluon or by a quark. As an example, if we plot, for 3-jet events from e^+e^- collisions, the frequency of the jet to originate from a quark and a gluon, respectively, as a function of $P_{g,max}$, we get the distributions shown in Fig. 2. The figure should be interpreted in the following way. If, in an event, the jet with the highest probability of being a gluon jet has the probability value $P_{g,max}$, then the probability of it to really originate from a gluon is given by the value of the gluon distribution (y_g) at $P_{g,max}$, divided by the summed values of the gluon and quark distributions ($y_g + y_q$) also at $P_{g,max}$. It is then clear that, for a certain P_{cut}, the efficiency and purity for identifying the gluon jet, and thereby also the quark jet(s), in events containing only one gluon, can be expressed as

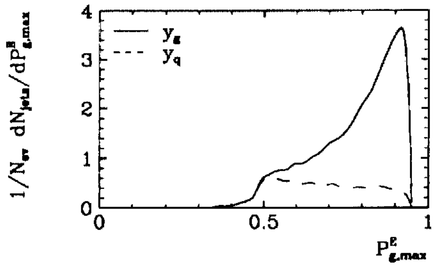

Figure 2: *The $P^B_{g,max}$ distributions for quarks and gluons in a 3-jet event.*

$$Efficiency = \frac{\int_{P_{cut}}^{1} (y_g + y_q)\, dP_{g,max}}{\int_{\frac{1}{\#jets}}^{1} (y_g + y_q)\, dP_{g,max}} \quad (3)$$

$$Purity = \frac{\int_{P_{cut}}^{1} y_g\, dP_{g,max}}{\int_{P_{cut}}^{1} (y_g + y_q)\, dP_{g,max}} \quad (4)$$

where $P_{cut} \geq \frac{1}{\#jets}$.

6 Identification of gluon jets using fragmentation properties

Due to the fact that gluons, according to QCD, carry a stronger colour charge than quarks, it is expected that there will be differences in their fragmentation. A large number of variables sensitive to these differences have been suggested for the purpose of performing quark-gluon separation. Above we have derived the probability formalism for a separation using the jet energies alone and here we will go through the same procedure for an identification by fragmentation variables, using a neural network.

Variables describing fragmentation properties are normally based on the relation between single particles in a jet and the jet axis. A jet algorithm therefore has to be applied before the fragmentation-sensitive variables can be calculated. To avoid a dependence of the fragmentation variables on the detailed reconstruction of a jet by different jet algorithms, we consider only particles in the jet core. The jet core is defined by taking the particles of a jet in descending order of P_l, the longitudinal momentum with respect to the jet axis, until we reach 80% of the total jet energy. Since the jet algorithm is not Lorentz-invariant, using only the jet core also leads to an insensitivity of the frame in which the clustering takes place, which is very important for the HERA events. We also want the fragmentation variables to be experimentally useful, i.e. they should not be greatly affected by detector smearing and poor event reconstruction.

One set of variables we have tested and found to provide the best separation between quarks and gluons is the so called **Fodor moments** [11]:

$$F_{nm}(E_{jet}) = \sum(\frac{P_T}{E_{jet}})^n\eta^m$$

where P_T and η are the transverse momentum and the pseudorapidity of a particle in a jet with respect to the jet-axis and E_{jet} is the jet energy. The sum is taken over all particles in the jet core. The three lowest moments have an obvious interpretation. F_{00} is the multiplicity of the jet, F_{01} is the pseudorapidity sum of all particles in the jet, and F_{10} is the transverse momentum sum of all particles, scaled by the jet energy. A careful study of the Fodor moments reveals that the results based on the different generators give general agreement only for some of the moments. This is, however, a necessary condition for obtaining an independence of both the jet production process and the fragmentation model used, and consequently we have concentrated on these moments. The mean values of the moments F_{11}, F_{1s}, F_{20} and F_{31} exhibit similar behaviour as a function of energy for all the generators except in the low energy range of the moment F_{1s} where the JETSET curves fall below the others. This is illustrated in Fig. 3a-d. A separation cut between quarks and gluons common for all the generators can thus, in principle, be found for the moments F_{11}, F_{20} and F_{31} over the full energy range, whereas this is not true for the moment F_{1s}. However, as can be observed in Fig. 3f, the Fodor moment distributions for quarks and gluons overlap significantly which in any case prevents a completely clean separation. As an example of a moment where the curves from the various generators are widely spread and therefore make the choice of a common separation cut difficult, we show the moment F_{00} in Fig. 3e. In our selection we have avoided the higher Fodor moments since they will, from an experimental point-of-view, be very sensitive to the reconstruction quality of the energy and direction of the particles.

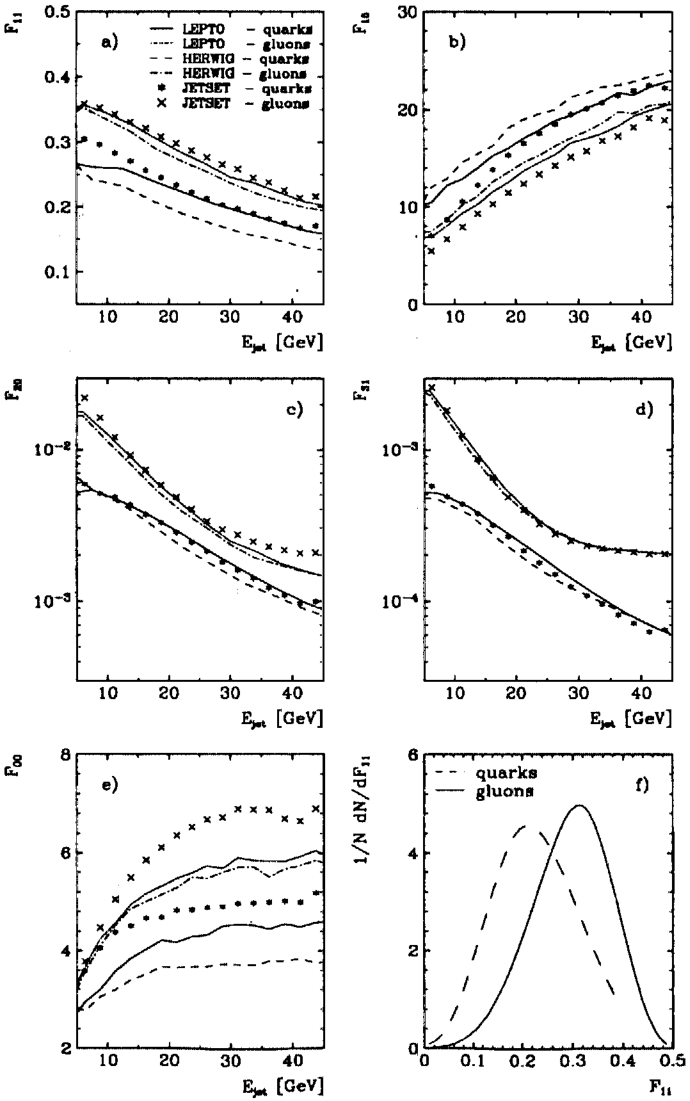

Figure 3: *The Fodor moments a) F_{11}, b) F_{1s}, c) F_{20}, d) F_{31} and e) F_{00} as a function of the jet energy. In f) the F_{11}-distributions for quarks and gluons in LEPTO at $E_{jet} \approx 25$ GeV are shown.*

One of the reasons for using a neural network in the jet identification based on fragmentation properties is that we want to simultaneously take into account the effect of several variables and their correlations. We have chosen a network, as implemented in the program package JETNET 2.0 [10], using the method of backpropagation, well suited for this kind of pattern recognition. We have varied the number of hidden layers, nodes and values of the learning rate, but this did not produce a significant change in the final results. We therefore decided to use one hidden layer and one output node ($0 = quark$; $1 = gluon$)

To enable a process-independent identification it is essential that the result does not depend on the jet energy, and one can therefore either try to select variables which are completely uncorrelated with the jet energy or train the neural network in such a way that the jet energy by itself does not give any discrimination. In the latter case it is important, from the point-of-view of the neural network, to be careful in the use of energy-dependent fragmentation variables. Although we want the network to be sensitive to the energy dependence of the fragmentation variables, it should not be affected by the jet energies themselves. Since the difference in the jet energy distributions is the most dominant effect, it might be picked up even implicitly by the neural network. Therefore a good training strategy will help to emphasize the learning of the network on the more subtle fragmentation properties.

Among the variables we have investigated it turns out that those most sensitive to differences in the fragmentation properties all have a considerable energy dependence (see Fig. 3). We thus trained the neural network with equal and flat energy distributions for quarks and gluons in order to prevent the network from being influenced by the jet energies themselves. Such an artificial training sample is obtained by using individual jets taken from the Monte Carlo-generated events. In the following we denote this method **balanced energy training.**

As input variables to the neural network we finally selected the Fodor moments F_{11}, F_{15}, F_{20} and F_{31} together with the jet energy, in order to provide the energy dependence of the fragmentation variables. Instead of using the notation $F_{nm,i}$ to specify the value of fragmentation variable F_{nm} for jet i, we will simplify the notation by letting F_i represent the values of all the fragmentation variables used for jet i. It has been proven, see e.g. [12], that the output of a neural network will simply be the conditional probability for a jet to be a gluon (or a quark, depending on how the output is defined), given the input variables and the composition of the training sample. Using Bayes' theorem, the probability for jet1 to be a gluon jet can be expressed through the density functions, given that the values of the fragmentation variables are F_1 at an energy $\tilde{E}_1$, in the following way:

$$P_g^F(jet1|F_1) = \frac{\tilde{f}_g^B(\tilde{E}_1)f_g^F(F_1)}{\tilde{f}_g^B(\tilde{E}_1)f_g^F(F_1) + \tilde{f}_q^B(\tilde{E}_1)f_q^F(F_1)} = \frac{f_g^F(F_1)}{f_g^F(F_1) + f_q^F(F_1)} \tag{5}$$

where $P_g^F(jet1|F_1)$ is thus identical to the network output. $\tilde{f}^B$ denotes the density function of the artificial energy distribution used in the balanced energy training, which implies $\tilde{f}_g^B(E_1) = \tilde{f}_q^B(E_1)$. Since the jet energy (E_1) is given as input to the network together with the fragmentation variables (F_1), the correct notation of the density function for the fragmentation should be $f^F(F_1|E_1)$. However, in the balanced energy sample the energy will provide information only on the energy dependence of the fragmentation variables, as already explained, which means that we are in reality considering only the fragmentation variables. In order to avoid confusion, we have therefore decided to use the simplified notation $f^F(F_1)$.

The probability (5) based on the fragmentation properties is also valid in the case where, instead of treating individual jets, we make use of the fact that we know the number of quark and gluon jets in the event. Thus it is not necessary to train the network specifically for this situation. For 2-jet events, the output of a network trained on the quark and the gluon in a pair according to the balanced energy method, is just a simple function of the output from a network trained on individual jets, assuming that the jets fragment independently. Since we are only considering the jet core, this assumption is justified. Given the energies E_1 and E_2, we consequently have $f_{gq}^F(F_1, F_2|E_1, E_2) = f_g^F(F_1|E_1)f_q^F(F_2|E_2) = f_g^F(F_1)f_q^F(F_2)$, using our simplified notation, and again, due to the balanced energy training, the combined density functions $\tilde{f}_{gq}^B(E_1, E_2) = \tilde{f}_{qg}^B(E_1, E_2)$.

$$P_{gq}^{F_{12}}(jet1, jet2|F_1, F_2) = \frac{f_g^F(F_1)f_q^F(F_2)}{f_g^F(F_1)f_q^F(F_2) + f_g^F(F_2)f_q^F(F_1)} \tag{6}$$

Dividing the nominator and the denominator by $[f_g^F(F_1) + f_q^F(F_1)][f_g^F(F_2) + f_q^F(F_2)]$ and using the fact that, for individual jets the quark and gluon probabilities are related as $P_q^F(jet1|F_1) = 1 - P_g^F(jet1|F_1)$, we get:

$$P_{gq}^{F_{12}}(jet1, jet2|F_1, F_2) = \frac{P_g^F(jet1|F_1)[1 - P_g^F(jet2|F_2)]}{P_g^F(jet1|F_1)[1 - P_g^F(jet2|F_2)] + [1 - P_g^F(jet1|F_1)]P_g^F(jet2|F_2)}$$

where $P_{gq}^{F_{12}}(jet1, jet2|F_1, F_2)$ is now expressed in single jet probabilities which are identical to the network output values.

The 3-jet events from e^+e^- collisions contain a quark and an anti-quark which are identical from the point-of-view of the fragmentation, and the probability for a gluon jet is obtained by a simple extension of the expression for the 2-jet events:

$$P_{gqq}^{F_{123}}(jet1, jet2, jet3|F_1, F_2, F_3) = \frac{f_g^{F_1} f_q^{F_2} f_q^{F_3}}{f_g^{F_1} f_q^{F_2} f_q^{F_3} + f_g^{F_2} f_q^{F_1} f_q^{F_3} + f_g^{F_3} f_q^{F_1} f_q^{F_2}}$$

with $f_g^{F_i} = f_g^F(F_i)$, i=1,2,3. The probability for a gluon jet in a 3-jet event can be expressed in terms of individual jet probabilites using a similar procedure as for the 2-jet case.

$$P_{gqq}^{F_{123}}(jet1, jet2, jet3|F_1, F_2, F_3) =$$
$$\frac{P_g^{F_1}(1 - P_g^{F_2})(1 - P_g^{F_3})}{P_g^{F_1}(1 - P_g^{F_2})(1 - P_g^{F_3}) + P_g^{F_2}(1 - P_g^{F_1})(1 - P_g^{F_3}) + P_g^{F_3}(1 - P_g^{F_1})(1 - P_g^{F_2})} \tag{7}$$

where $P_g^{F_i} = P_g^F(jeti|F_i)$, i=1,2,3.

7 Jet identification using jet energy and fragmentation properties

We now want to extract the conditional probability for a jet being a gluon jet, given both the jet energies and the values of the fragmentation variables which were used as input to our neural network. In the (2+1) jet case we obtain this by using the equations (2) and (6):

$$P_{gq}^{C_{12}}(jet1, jet2|E_1, E_2, F_1, F_2) = \frac{f_{gq}^{B_{12}} f_g^{F_1} f_q^{F_2}}{f_{gq}^{B_{12}} f_g^{F_1} f_q^{F_2} + f_{gq}^{B_{21}} f_g^{F_2} f_q^{F_1}}$$

where

$$f_{gq}^{B_{ij}} = \frac{f_{gq}^{B}(E_i, E_j)}{f_{gq}^{B}(E_i, E_j) + f_{gq}^{B}(E_j, E_i)}$$

If we now divide all the terms by $(f_{gq}^{B_{12}} + f_{gq}^{B_{21}})(f_g^{F_1} + f_g^{F_1})(f_q^{F_2} + f_q^{F_2})$ we obtain

$$P_{gq}^{C_{12}}(jet1, jet2|E_1, E_2, F_1, F_2) =$$
$$\frac{P_{gq}^{B_{12}} P_g^{F_1} P_q^{F_2}}{P_{gq}^{B_{12}} P_g^{F_1} P_q^{F_2} + P_{qg}^{B_{12}} P_q^{F_1} P_g^{F_2}} = \frac{P_{gq}^{B_{12}} P_g^{F_1}(1 - P_g^{F_2})}{P_{gq}^{B_{12}} P_g^{F_1}(1 - P_g^{F_2}) + (1 - P_{gq}^{B_{12}})(1 - P_g^{F_1}) P_g^{F_2}}$$

After division of all the terms by $P_g^{F_1}(1 - P_g^{F_2}) + (1 - P_g^{F_1}) P_g^{F_2}$ we finally get

$$P_{gq}^{C_{12}}(jet1, jet2|E_1, E_2, F_1, F_2) = \frac{P_{gq}^{B_{12}} P_{gq}^{F_{12}}}{P_{gq}^{B_{12}} P_{gq}^{F_{12}} + P_{gq}^{B_{21}} P_{gq}^{F_{21}}} \tag{8}$$

where we have used $P_{gq}^{B_{12}} = 1 - P_{gq}^{B_{21}}$ and $P_{gq}^{F_{12}} = 1 - P_{gq}^{F_{21}}$.

In a completely analogous way the corresponding conditional probability can be obtained for a 3-jet event.

$$P_{gqq}^{C_{123}}(jet1, jet2, jet3|E_1, E_2, E_3, F_1, F_2, F_3) =$$
$$\frac{P_{gqq}^{B_{123}} P_g^{F_1}(1 - P_g^{F_2})(1 - P_g^{F_3})}{P_{gqq}^{B_{123}} P_g^{F_1}(1 - P_g^{F_2})(1 - P_g^{F_3}) + P_{gqq}^{B_{312}}(1 - P_g^{F_1})(1 - P_g^{F_2}) P_g^{F_3} + P_{gqq}^{B_{231}}(1 - P_g^{F_1}) P_g^{F_2}(1 - P_g^{F_3})} =$$
$$\frac{P_{gqq}^{B_{123}} P_{gqq}^{F_{123}}}{P_{gqq}^{B_{123}} P_{gqq}^{F_{123}} + P_{gqq}^{B_{312}} P_{gqq}^{F_{312}} + P_{gqq}^{B_{231}} P_{gqq}^{F_{231}}} \tag{9}$$

8 Results

Fig. 4a shows the neural network output for individual jets in 3-jet events from e^+e^- collisions, based on the fragmentation variables discussed in section 6 and their energy dependence. The x-axis gives directly the probability that a jet from a quark and a gluon, respectively, is identified as a gluon jet. Using expression (7) we can calculate the event-based conditional probability of a jet to originate from a gluon, given the values F_1, F_2, F_3 for the fragmentation variables and the energies $\tilde{E}_1, \tilde{E}_2, \tilde{E}_3$, specified to account for the energy dependence of the fragmentation variables. From these calculations we select for each event the jet with the highest probability of being a gluon jet, giving a distribution as shown in Fig. 4b. We note that the distributions populate exactly the allowed region $1/3 < P_{g,max} < 1$.

Fig. 4c presents the event-based probability of a jet being a gluon jet as obtained from the ME calculation according to equation (1), and Fig. 4d gives the distributions of jets with the highest conditional probability, in the event, of coming from a gluon. In agreement with what has been indicated previously in the text, Figs. 4b and d confirm that the jet energies are much more efficient in identifying jets than are the fragmentation properties.

The event-based combined conditional probability for gluon jet identification, given the jet energies E_1, E_2, E_3 and the values F_1, F_2, F_3 for the fragmentation variables, achieved using equation (9), is given in Fig. 4e and finally, the jet per event with the highest combined probability of being produced by a gluon is plotted in Fig. 4f. The corresponding plots for (2+1) jet events from ep collisions are shown in Figs. 5a-f.

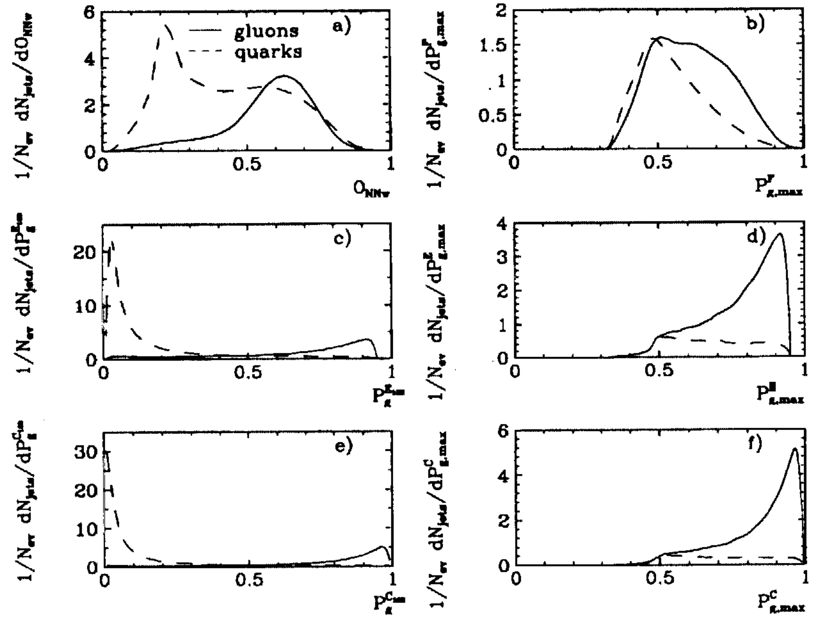

Figure 4: *Results for 3-jet events from e^+e^- shown as, a) the neural network output for individual jets, b) the $P^F_{g,max}$ distribution from the fragmentation, c) the gluon probability obtained from the ME calculation, d) the $P^E_{g,max}$ distribution from the jet energies, e) the combined conditional probability distribution according to equation (9), f) the $P^C_{g,max}$ distribution from the combined conditional probability.*

From the $P_{g,max}$ distributions we can now calculate the purity and efficiency as a function of P_{cut} according to expressions (3) and (4). In Fig. 6 the purity is plotted versus the efficiency separately for an identification based on the jet energies and on the fragmentation variables as well as for a combination of the two. Fig. 6a shows the results for 3-jet events generated with JETSET and for which the neural network has also been trained on a sample generated with JETSET. The corresponding results for (2+1) jet events are shown in Figs. 6b and c for samples generated with LEPTO and HERWIG, respectively, using networks trained on jets from the same generators. In this context, one should bear in mind that the *a*

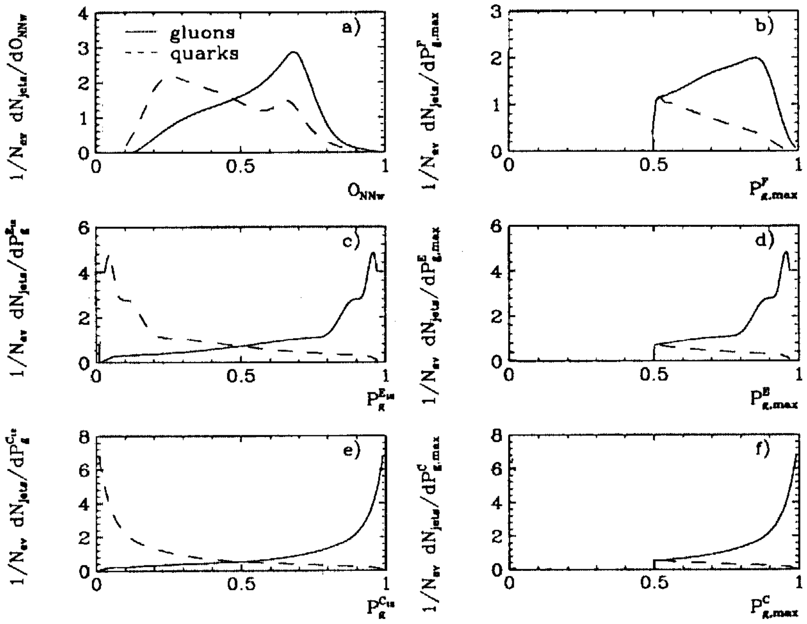

Figure 5: *Results on (2+1)-jet events from ep collisions shown as, a) the neural network output for individual jets, b) the $P^F_{g,max}$ distribution from the fragmentation, c) the gluon probability distribution according to equation (2), d) the $P^E_{g,max}$ distribution from the jet energies, e) the combined conditional probability distribution according to equation (8), f) the $P^C_{g,max}$ distribution from the combined conditional probability.*

priori probability of a correct gluon identification is 33% for 3-jet events, and 50% for 2-jet events. This is also the reason why the vertical axes of Fig. 6 have different starting points. **It must also be stressed that the final results presented here correspond to an identification of all jets in an event and they can therefore not be directly compared with results from identification of individual jets.**

The identification using fragmentation variables in general gives much poorer separation between quarks and gluons than the energy-based identification. However, the methods are normally complementary in the sense that events which are well separated by the jet energies are not necessarily those which are well separated by the fragmentation properties. Consequently an improved result is expected if the identifications from energy and fragmentation are combined.

From a comparison of Figs. 6a-c it can be seen that the energy-based identification of 3-jet events at LEP and (2+1) jet events at HERA are both 78-79% at 100% efficiency, but as an increasingly harder P_{cut} is made in the $P_{g,max}$ distributions the identification improves faster for the *ep* events than for the e^+e^- events. Concerning the fragmentation-based identification, it seems to give significantly better results for the *ep* events than for the e^+e^- events over the full efficiency range. This is in particular true for the HERWIG sample, for which the fragmentation-based identification is almost equally good as the one based on jet energies. The combined probabilities give improved identifications with respect to the jet energy results, which are essentially equal for the JETSET and LEPTO samples ($\approx 5\%$), while the improvment is somewhat greater for the HERWIG sample ($\approx 9\%$).

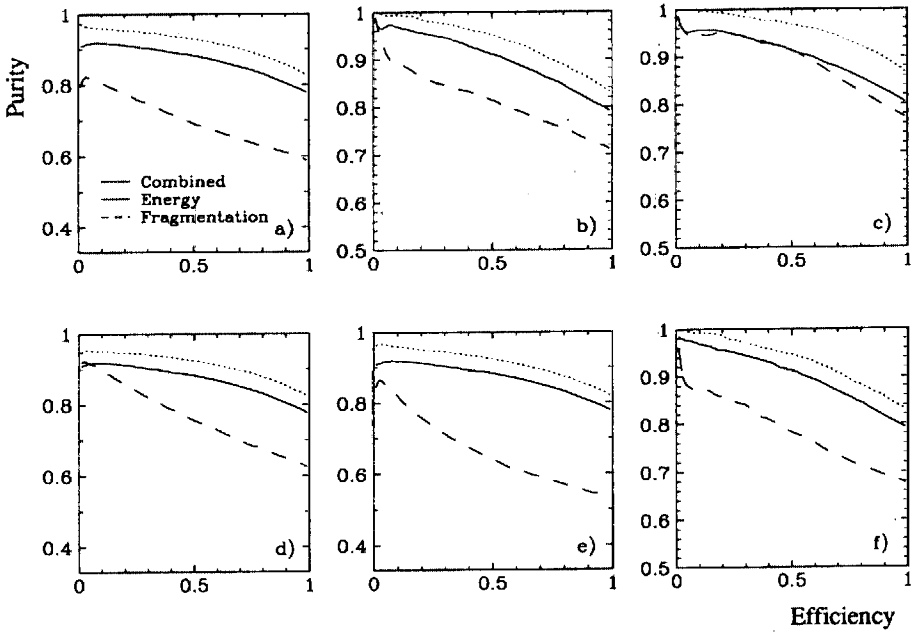

Figure 6: *The purity as a function of the efficiency for a network a) trained on a JETSET sample and tested on a JETSET sample, b) trained on a LEPTO sample and tested on a LEPTO sample, c) trained on a HERWIG sample and tested on a HERWIG sample, d) trained on a LEPTO sample and tested on a JETSET sample, e) trained on a HERWIG sample and tested on a JETSET sample, f) trained on a HERWIG sample and tested on a LEPTO sample.*

In order to investigate the process- and model-dependence we also present results from samples using a generator different from the one used in the training of the neural network.

These results are given in Figs. 6d-f. A comparison between Figs. 6a and d illustrates the process-dependence, while a comparison of Figs. 6b and f provides the model-dependence. From Figs. 6a and e the effects of both the model- and process-dependence can be extracted.

According to Figs. 6a and d the fragmentation result is better when a test sample from JETSET is presented to a network trained on LEPTO compared with a network trained on JETSET. The explanation of this strange behaviour is that the energy dependence of some Fodor moments are different for LEPTO and JETSET, causing the network to implicitly pick up this energy dependence in spite of the balanced energy training. Although it, in principle, should be possible to obtain a process-independent identification of jets using fragmentation variables, our results indicate that this is difficult to achieve in practice. The reason for this is most certainly that the phenomenological models used to generate e^+e^- events and ep events do not give a perfect description of these processes.

Concerning the model dependence of the fragmentation-based identification we have observed that a network trained to a HERWIG sample and applied on a HERWIG test sample gives a much higher degree of identification than a network trained on a LEPTO sample and tested on a LEPTO sample. This difference must be due to the different fragmentation models used in HERWIG and LEPTO. However, if we now compare Figs. 6b and f we notice that the curves are almost identical, indicating that the separation between quarks and gluons is optimized in the same way by the two networks trained on LEPTO and HERWIG samples. Thus a common network can be used to differentiate between quarks and gluons for samples generated with both LEPTO and HERWIG.

9 Conclusions

We have studied the problem of identifying jets using the jet energies and fragmentation variables separately. Jets produced in simulated e^+e^- and ep collisions were used to investigate a possible process-independent identification, and two different fragmentation schemes were used to study the model dependence.

The conditional probability of a jet to originate from a gluon (or a quark) can be calculated from Bayes' theorem provided the density functions for gluons and quarks with respect to jet energies and fragmentation variables are known. The formalism for extracting these probabilities, for the event types investigated here, has been presented. The advantages of working with probabilities are the simple interpretation of the results and the procedure of combining the results from the identifications based on the jet energy and the fragmentation variables. Since jet identification based on fragmentation variables is a multidimensional problem, the neural network technique is the only feasible way to convert the full information on the fragmentation into a probability of having a certain type of jet. By using the neural network as an estimate of Bayesian probabilities, all calculations are performed in the firm framework of mathematical statistics, which is clearly advantageous compared with using the neural network as a black-box classifier.

The event-based level of identification using the jet energies is about 80% at 100% efficiency for both e^+e^- 3-jet events and ep QCD-Compton events, whereas the fragmentation-based identification gives 60% purity for the e^+e^- events and more than 70% for the ep events, also at 100% efficiency. A combination of the two leads to improvments, over the full efficiency range, of between 5% and 9% with respect to the results from the jet energy alone. The identification of the jets in a complete event is a much stronger requirement than the identification of individual jets and therefore a direct comparison of such results is not possible.

One advantage of performing the jet identification based on jet energies and fragmentation variables separately is that the neural network will concentrate on extracting the subtle differences in the fragmentation of quarks and gluons without being influenced by the differences in the jet energies. Another advantage is that it allows an investigation of whether a process independent identification is possible, using the fragmentation properties. The fragmentation-based results are observed to be significantly different depending on the fragmentation model used. On the other hand, the variation of the results from a combination of the energy- and fragmentation-based identifications is much less striking, which is consistent with previous observations.

Acknowledgements: The analysis described in this paper as well as the way the results should be presented have gone through several metamorphoses as the work has proceeded. In the initial phase we had many animated discussions with O. Bärring and V. Hedberg, which forced us to find convincing arguments for our analysis strategy. We are indebted to G. Ingelman and M. Lindström for their careful reading of the manuscript and for giving constructive comments, suggestions and critisism. Three of us would like to thank the DESY Directorate for its kind hospitality.

A Appendix

Bayes' Theorem

If the entire event space is composed of the subsets $B_i, (i = 1...n)$, with no elements in common, then the subsets are said to be *mutually exclusive* and *exhaustive*, which means that

$$\sum_{i=1}^{n} P(B_i) = 1 \tag{10}$$

Provided A is also a set that belongs to the event space, Bayes' theorem states

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} \tag{11}$$

This theorem can be proven by starting from the definition of the conditional probability, $P(A \cap B) = P(B|A)P(A) = P(A|B)P(A)$, where $P(B|A)$ is to be interpreted as the probability that the event B occurs under the condition that A has already occurred. For our subset B_i we thus get

$$P(A \cap B_i) = P(B_i|A)P(A) = P(A|B_i)P(B_i)$$

$$\Rightarrow P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} \tag{12}$$

The elements of a set might be classified according to more than one criterion so that, for example, we have $\sum_{i=1}^{m} P(A_i) = \sum_{j=1}^{n} P(B_j) = 1$. If some of the criteria are being neglected in the classification we can define the *marginal probability* for A_i according to

$$P(A_i) = \sum_{j=1}^{n} P(A_i \cap B_j) = \sum_{j=1}^{n} P(A_i|B_j)P(B_j) \tag{13}$$

where we have again used the definition of the conditional probability for the second step. Using this expression we can now rewrite equation (12) to obtain equation (11).

If we now consider the (2+1) jet case, we could take the subset B_i to represent the four configurations of quark- and gluon jets possible if the jets are identified individually in the event. These are $B_i = \{gq, qg, qq, gg\}$, where, for example, gq means that jet1 is a gluon jet and jet2 a quark jet. Since we have selected events which have one quark and one gluon in the final state (QCD-Compton events) we introduce this information into our probabilities by considering only the allowed configurations, and define the two subsets $B_1 = \{gq\}$, $B_2 = \{qg\}$. Assuming that the two jets are in a state C, which might refer to energy and/or fragmentation variable values, we can compute the probability of having a gq configuration in the state C, by using Bayes' theorem.

$$P(gq|C) = \frac{P(gq)P(C|gq)}{P(gq)P(C|gq) + P(qg)P(C|qg)} \tag{14}$$

This formula is valid when state C has a positive probability $P(C) > 0$. When C is defined in terms of a continuously varying quantity, the discrete probabilities in Bayes formula should

be replaced by probability density functions. For example, if C represents a continuous energy variable E, then (14) reads

$$P(gq|E) = \frac{P(gq)f_{gq}(E)}{P(gq)f_{gq}(E) + P(qg)f_{qg}(E)} \tag{15}$$

where f_{gq} and f_{qg} are the joint energy density functions for pairs gq and qg, respectively.

For the 3-jet case, we have 8 possible configurations of quarks and gluons. After deleting the impossible combinations, the remaining subsets are $B_i = \{gqq, qgq, qqg\}$. The final expression becomes

$$P(gqq|C) = \frac{P(gqq)P(C|gqq)}{P(gqq)P(C|gqq) + P(qgq)P(C|qgq) + P(qqg)P(C|qqg)} \tag{16}$$

References

[1] L. Lönnblad, Peterson and Rögnvaldsson, Phys. Rev. Lett. 65(1990)1321.
L. Lönnblad, Peterson and Rögnvaldsson, Nucl. Phys. B349(1991)675.
Csabai, Czako and Fodor, Nucl. Phys. B374(1992)288.

[2] T. Sjöstrand, Computor Phys. Comm. 39 (1986) 347, it ibid 43 (1987) 367.

[3] B. Andersson, G. Gustafson, G. Ingelman, T. Sjöstrand, Phys. Rep. 97(1983)31.

[4] G. Ingelman, LEPTO version 6.1, in Proceedings 'Physics at HERA', Eds W. Buchmüller, G. Ingelman, DESY Hamburg 1992, vol. 1-3 and references therein.

[5] G. Marchesini, B.R. Webber, G. Abbiendi, I.G. Knowles, M.H. Seymour and L. Stanco, Computer Phys. Comm. 67(1992)465.

[6] T. Sjöstrand, Computer Phys. Comm. 28 (1983) 227

[7] V. Hedberg, G. Ingelman, C. Jacobsson, L. Jönsson, Z. Phys. C63 (1994) 49.

[8] D. Graudenz, N. Magnussen,
M. Fleischer et al.,
V. Hedberg, G. Ingelman, C. Jacobsson, L. Jönsson,
Proceedings 'Physics at HERA', Eds W. Buchmüller, G. Ingelman, DESY Hamburg 1992, vol. 1, p. 261, 303, 331.

[9] K. C. Hoeger, Measurement of x, y, Q^2 in Neutral Current Events, in Proceedings 'Physics at HERA', Eds W. Buchmüller, G. Ingelman, DESY Hamburg 1992, vol. 1, p. 43

[10] L. Lönnblad, C.Peterson, H. Pi and T. Rögnvaldsson, Comp. Phys. Comm. 67(1991)193.

[11] Fodor, Phys. Rev. D41(1990)1726.

[12] M.D. Richard, R.P. Lippman, Neural Comp 3 (1991) 461.