ALEPH 95-006
PHYSIC 95-006
24 January 1995

Status of an analysis of $Z \rightarrow b\bar{b}\, b\bar{b}$ using a neural network

G. Hansper
IHEP Heidelberg

Abstract

This note reports on the status of an analysis searching for events containing four b quarks. 32219 preselected events from '92/ '93 data and Monte Carlo are processed using a feed forward neural network.
Currently the b–tagging efficiency is not sufficient to discriminate signal from background events. A systematic deviation in the tagging variables is observed between data and Monte Carlo.

1 Introduction

The ALEPH vertex detector allows efficient tagging of Z decays into a $b\bar{b}$ pair by measuring the impact parameters of charged tracks with respect to the Z decay point. The tagging of primary b quarks is generally performed by the QIPBTAG package [1]. In this analysis QIPBTAG has been used to try to also tag secondary $b\bar{b}$ pairs in 4–b events[1] , e.g. those resulting from gluon splitting in the reaction

$$Z \rightarrow b\bar{b} + g \rightarrow b\bar{b} + b\bar{b} \; .$$

A determination of the cross section for this process would yield information about the mass effect in gluon splitting into heavy quarks. Furthermore, these processes could give a contribution in the measurement of $\Gamma_{b\bar{b}}$, and final states with four b quarks might become an interesting channel in the search for the Higgs boson at LEP II.

This note is organized as follows: In the next section the jet finding algorithm and the expected number of events are discussed. Then the procedure for preselecting events is described, followed by a presentation of the properties of the remaining events. Finally the results of the neural net analysis are discussed.

2 Expected number of events

In [2] the cross sections for the decay of the Z to four fermions are calculated at tree level to $O(\alpha_s{}^2)$, including the fermions' masses. The results are infrared divergent for massless fermions. In order to get a measurable quantity the four final state partons have been clustered to parton jets using the JADE algorithm with a fixed $y_{cut} = 0.01$. A 4–b event is obtained if the four final state partons are all b quarks *and* the jet clustering produces exactly four parton jets. This calculation yields one 4–b event in 10000 hadronic Z decays.
The JETSET 7.3 [3] parton shower Monte Carlo predicts one 4–b event in 3500 hadronic Z decays. This result is obtained by simply counting the number of b quarks in the parton final state. When using the same jet clustering as above this also reduces to one event in 10000 hadronic Z decays where four jets on parton level are present with each jet containing a b quark. (Since the parton shower model in general produces more than four final state partons one has to make sure that in a four–jet event each jet contains a b quark.)
Assuming $4 \cdot 10^6$ Z's by the end of LEP I running these numbers amount to $\sim$ 800 4–b events, where one can hope to observe about 300 of these as four–jet events.

A more efficient method of jet clustering is to force all events into a four–jet geometry by varying the y_{cut}. The JADE and DURHAM algorithms have been compared in their ability to cluster 4–b events into four b–jets. Here the DURHAM–P scheme proved to have the best performance: $\sim$ 70% of the 4–b events have been clustered into four parton jets with each jet containing a b quark. This algorithm has also been chosen for the further analysis, working on energy flow (EFLW) objects.
In order to reject clear two and three jet events that have only artificially been divided into four jets, a cut has been made at detector level requiring that each pair of jets makes a solid angle of at least 40°. These jets have been related to parton level jets by matching those which are closest in angle. In 85% of events this procedure is free of ambiguities, i.e. there is a one–to–one correspondence between jets on parton and detector level.

[1] In the following, "b quark" will always refer to the b as well as the $\bar{b}$. The events will be called "4–b events".

Taking into account the 40° cut we again end up with ~ 300 'reconstructed' events.

3 Preselection of event samples

Three samples have been used for this analysis:

3565	4–b MC events (generated with JETSET 7.3),
1859870	hadronic MC events ('92 production),
1291162	data events ('92 and '93 with CLASS16) .

Some standard quality cuts have been applied to these before further processing.

$$\left.\begin{array}{lcl} D0 & \leq & 2 \quad cm \\ Z0 & \leq & 5 \quad cm \\ \text{Number of TPC hits} & > & 4 \\ \cos\theta & < & 0.94 \\ 200 \quad MeV/c \leq \quad P & \leq & 55 \quad GeV/c \end{array}\right\} \quad \text{charged track selection}$$

and :

$$\left.\begin{array}{lcl} Nch & \geq & 6 \\ Ech & \geq & 15 \quad GeV \\ \cos\theta_{Thrust} & \leq & 0.95 \\ 45 \quad GeV \leq Evis & \leq & 137 \quad GeV \\ \frac{\Sigma P_z}{Evis} & < & 0.4 \end{array}\right\} \quad \text{event selection (EFLW objects)}$$

The jetfinding algorithm described above has been performed including the requirement that each jet axis lie in the acceptance of the vertex detector ($|cos(\Theta_{jet,beam})| \leq 0.85$).
The jets are ordered by energy; in the following a subscript i on a jet variable refers to the i^{th} jet in the energy–ordered list.
Then the probability of each jet to have originated from the primary vertex is calculated by using QIPBTAG. The standard QIPBTAG track selection in ALPHA 117 has been used with one difference: a cut that requires a track to lie close ($|cos(\Theta_{jet,track})| \geq 0.7$) to one jet axis has been dropped. The reason is that the jets are not expected to be highly collimated. The QIPBTAG track probabilities are then combined to give the corresponding jet probabilities. A soft cut is made requiring that each jetprobability P_i be less than 0.9 . This assures that each jet posesses at least one charged track with lifetime information.
Here, we find a systematic discrepancy between Monte Carlo events and real data: The data events show higher impact parameters (i.e. more lifetime) than the Monte Carlo events. This can be seen in figure 9, where the ratio of the normalized distributions in the jet probabilities is shown. A clear deviation from a flat distribution is observed.
That this discrepancy is not only a feature of the special event topology has been checked by running QIPBTAG on '92 data and Monte Carlo. Only those events in which QIPBTAG has found two jets have been used to compare the jet probabilities. Similar discrepancies between data and MC are found.

Cut	4–b sample	hadronic MC sample	Data sample
Track – / event – selection	(98.2± 1.6)%	(95.22± 0.07)%	(97.50± 0.09)%
Jetfinding	(27.5± 0.9)%	(6.95± 0.02)%	(7.38± 0.02)%
Lifetime $P_i < 0.9$	(21.1± 0.8)%	(2.72± 0.01)%	(2.89± 0.01)%
Jetmultiplicity $N_{EFLW} \geq 3$	(20.0± 0.7)%	(2.28± 0.01)%	(2.50± 0.01)%

Table 1: The fraction of events passing the preselection cuts for the three different samples.

A final preselection of the events is done by requiring a minimum multiplicity of at least three EFLW objects in each jet.
The fractions of events surviving these cuts are given in table 1. The 4–b events that are left in the MC hadronic sample are removed and for the further analysis we are left with

712	4–b events,
42415	MC background events ,
32219	data events.

4 Properties of the preselected events

In this section the properties of the remaining 4–b events that will be used for the neural net analysis are presented.
One of the dominant features results from the fact that the gluons splitting into $b\bar{b}$ are rather hard. The lower bound on the gluon energy is given by $10\,GeV$ (twice the b mass) with the distribution having a mean value of $33\,GeV$. This is reflected in the distribution of jetenergies and minimal invariant mass of the jet pairs shown in figures 1 and 2. It should be noted that the two lowest–energy jets do not necessarily correspond to the secondary b quarks. Though the highest–energy jet is likely to correspond to one primary b quark, in general the quark radiating the gluon loses a large amount of energy.
From the parton level kinematics one would naively expect to have the highest–energy jet in one hemisphere while the other hemisphere should contain the three other jets. That this is only approximately true can be seen in figure 3. Besides the relatively bad angular resolution of the jet axes, gluon radiation of the highest energy quark can modify this simplistic picture even on parton level.
The main background to 4–b events will be $q\bar{q}\ q\bar{q}$ or $q\bar{q}\ b\bar{b}$ events, respectively. Since B hadron decays result in jets with high transverse momentum these background events may partly be rejected by requiring high values of sphericity and aplanarity. These eventshape variables are shown in figure 4.
While the above variables are sensitive to the topology and kinematics of the events, the central point will be the b–tagging. Here, the number of high–momentum ($p > 2\,GeV/c$) leptons can be used as an indicator for 4–b events (figure 5), but the more powerful variables are the QIPBTAG probabilities of which the sum of the logarithms of the single jet probabilities is shown in figure 6.
One last variable used for the neural net analysis is the number of tracks QIBPTAG has used for calculating the jet probabilities (figure 7). This is intended to estimate the significance of the single jet probabilities: the more 'good' tracks a jet contains the more weight its probability can be given.

5 Neural Net Analysis

We have used JETNET 3.0 [4] to further process the preselected data. The net has been set up with 13 input nodes, one hidden layer with 7 nodes, and one output node. The learning algorithm is the standard Back Propagation Algorithm. The following variables have been fed into the input nodes:

- The four jet probabilities, $P_1 - P_4$
- The number of tracks used to calculate these probabilities, $NGT_1 - NGT_4$
- The Aplanarity and Sphericity of the event
- The number of reconstructed leptons with $p > 2\,GeV/c$
- The minimum invariant mass of two jets
- The cosine of the angle between the highest–energy jet and its nearest neighbour

One half of the 4–b sample and the background sample have been used for training the net with the aim that the value of the output node be '0' for 4–b events and '1' for background. Afterwards the net is being tested with the other half of each sample and the data events are fed into the net, too. A typical distribution of the output node's value is shown in figure 8 where the histograms have been normalized to 1. The separation performance of the net is given by cutting the distribution at a desired 4–b reconstruction efficiency and then calculating the signal over background (S / B) by counting the number of background events between '0' and the cut value. The corresponding numbers are given in table 2 together with the numbers for the data events. (The efficiencies refer to the original number of events in each sample, the S / B is determined by assuming S / B = 1/3500 without any cuts.) The

4–b efficiency	MC background	S / B	Data
5 %	(0.021±0.001)%	1 / 14	(0.028±0.002)%
10 %	(0.060±0.003)%	1 / 21	(0.086±0.003)%
15 %	(0.219±0.005)%	1 / 51	(0.285±0.005)%

Table 2: Fraction of MC / data events passing the neural net for different 4–b reconstruction efficiencies (errors are statistical only).

background is suppressed by a factor of roughly 200 with respect to the signal which is not sufficient to allow the extraction of a signal from the number of data events. Nevertheless one sees a significant excess in the fraction of data events compared to the Monte Carlo samples. This is largely explained by the discrepancy between the data and Monte Carlo samples found in the QIPBTAG variables. As mentioned above, data events show more lifetime than the Monto Carlo events. Since 'high lifetime' signals a 'good candidate' the neural net tends to assign lower output node values to these events, thus enhancing the data fraction.

Any similar analysis relying strongly on precise lifetime information will have to overcome these systematic deviations. Additionally, QIPBTAG has been optimized to tag primary B hadrons in two– or three–jet events; some work would have to be invested to improve its performance on four–jet events. Currently, this analysis hasn't entered a stage where systematics are the main problem. The principal question is whether it will be possible to further reduce the background. Looking at the composition of the Monte Carlo background at 10 % 4–b efficiency one finds:

- Most of the events contain a $b\overline{b}$ pair.
- 30 % are of the type $q\overline{q}$ $q\overline{q}$,
- 1.5 % have a secondary $b\overline{b}$ pair.
- The rest is dominated by hard ($\geq 20GeV$) gluon emission with the distribution of jet energies similar to 4–b events.

Some methods have been tried to reject more of this background (variations of the tagging variables, jetshape variables), without success. Perhaps sophisticated quark– / gluon–jet discrimination might help to reduce the $b\overline{b}$ g background, but it seems to be difficult to get rid of the $b\overline{b}$ $q\overline{q}$ events that show lifetime in each of the jets.

6 Conclusions

An analysis has been presented that searches for 4–b quark final states using a neural network. With the current variables and methods it does not seem possible to find a significant signal. The b–tagging efficiency for four–jet events would have to be greatly improved in order to remove background from gluon splitting to light quarks.
Nevertheless, estimations show that the 4–b channel might become one possibility to look for a Higgs signal at LEP II.

Acknowledgements

I would like to thank H.Stenzel and M.Wunsch for their interest, ideas, and support.

Figures

All histograms have been normalized to yield 1 when summing the bin contents.

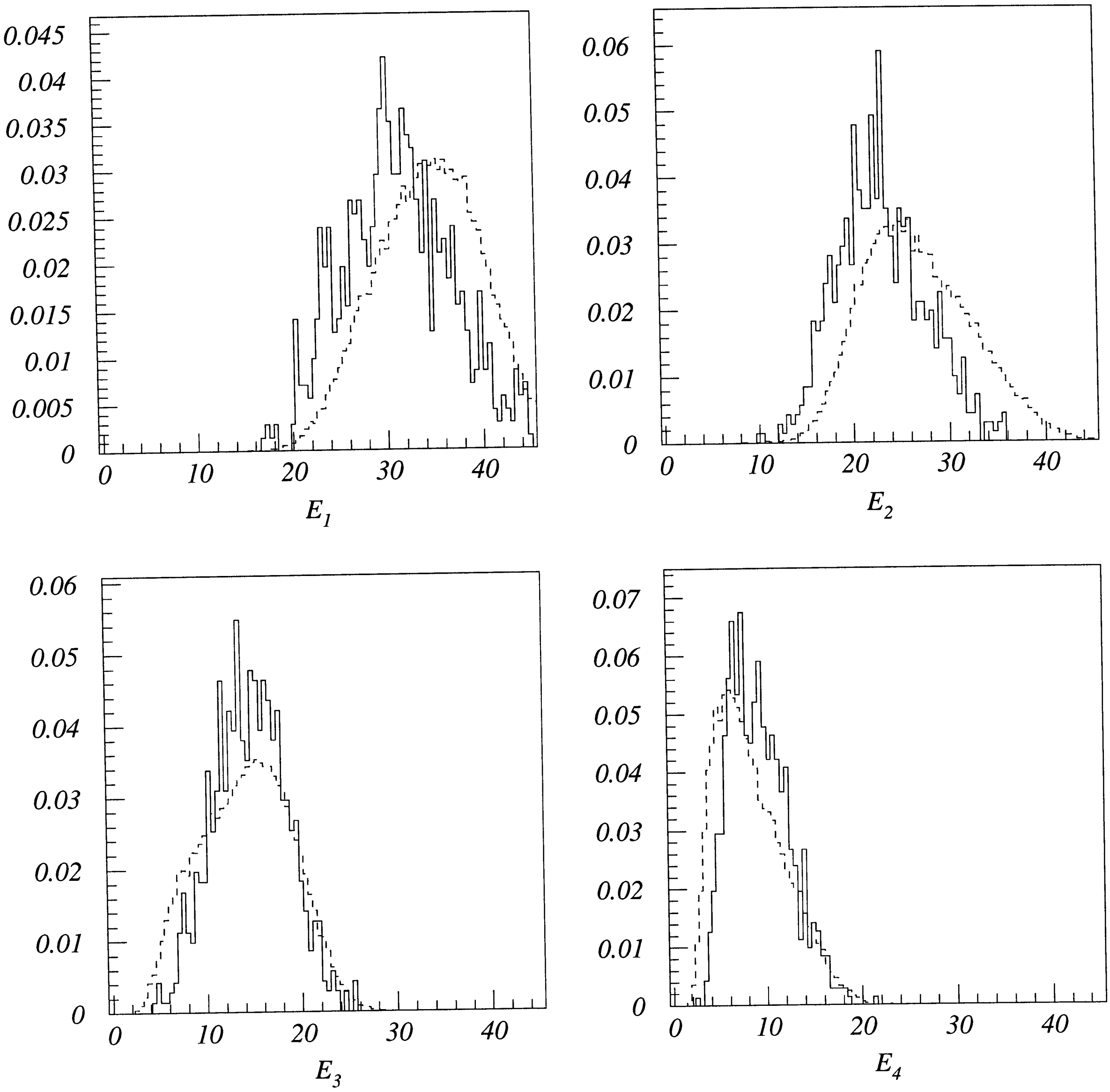


Figure 1: The normalized distributions of the jet energies $[GeV]$ for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.

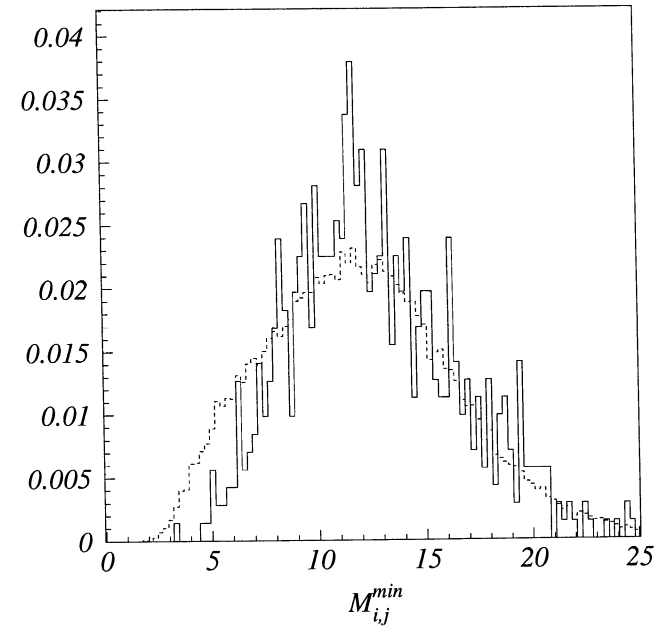


Figure 2: The normalized distributions of the minimum invariant mass $[GeV]$ of two jets for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.

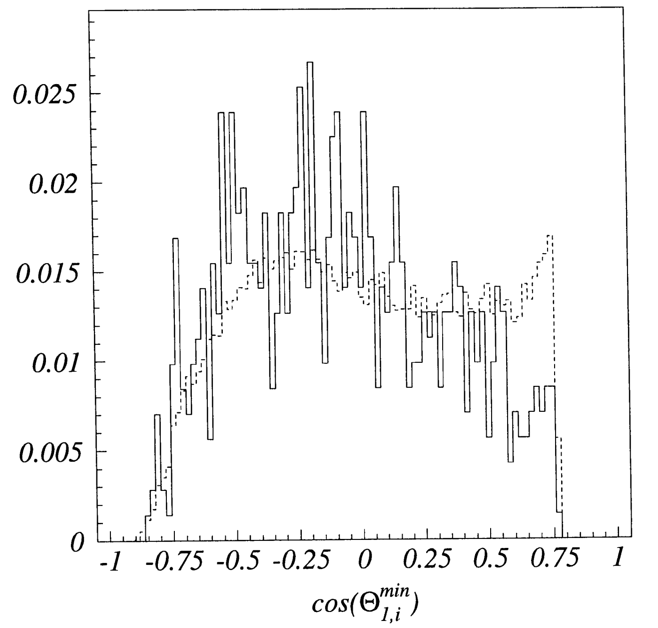


Figure 3: The normalized distributions of the cosine of the angle between the highest–energy jet and its nearest neighbour for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.

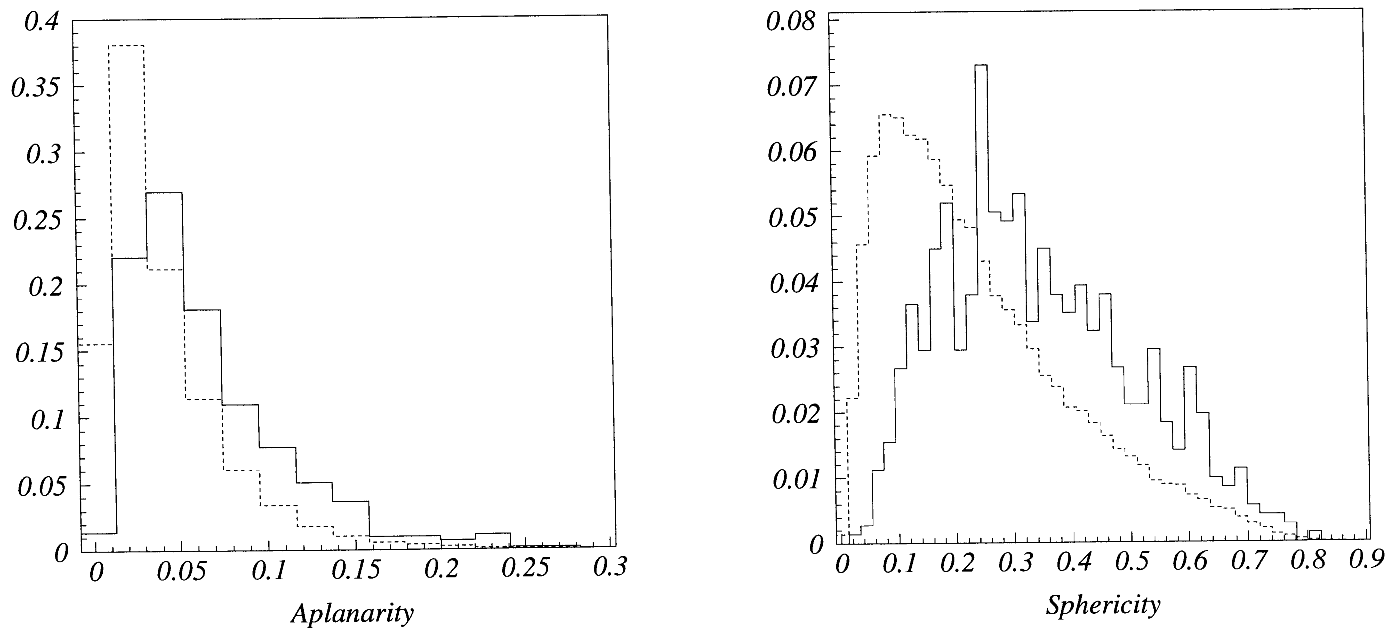
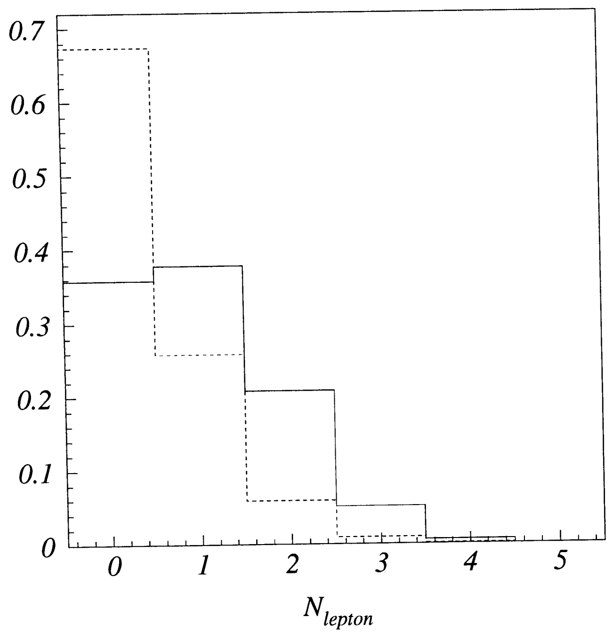


Figure 4: The normalized distributions of the Aplanarity and Sphericity for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.



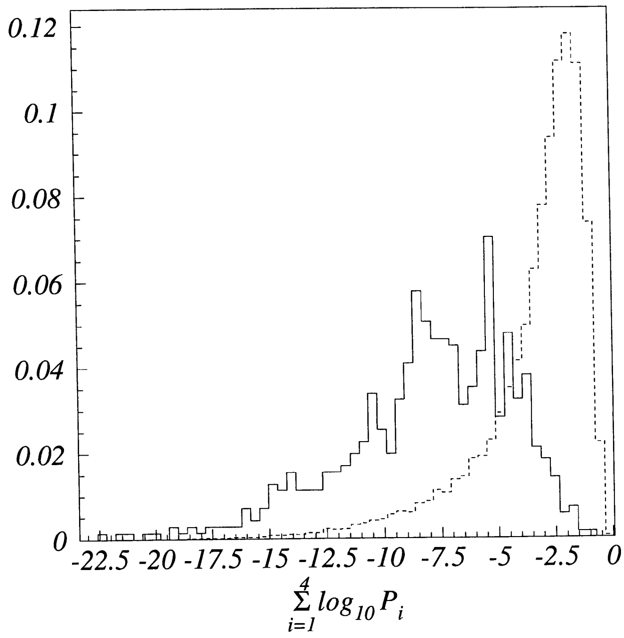


Figure 5: The normalized distributions of the number of reconstructed leptons with $p > 2\,GeV$ for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.

Figure 6: The normalized distributions of the sum of logarithms of the QIPBTAG jet probabilities for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.

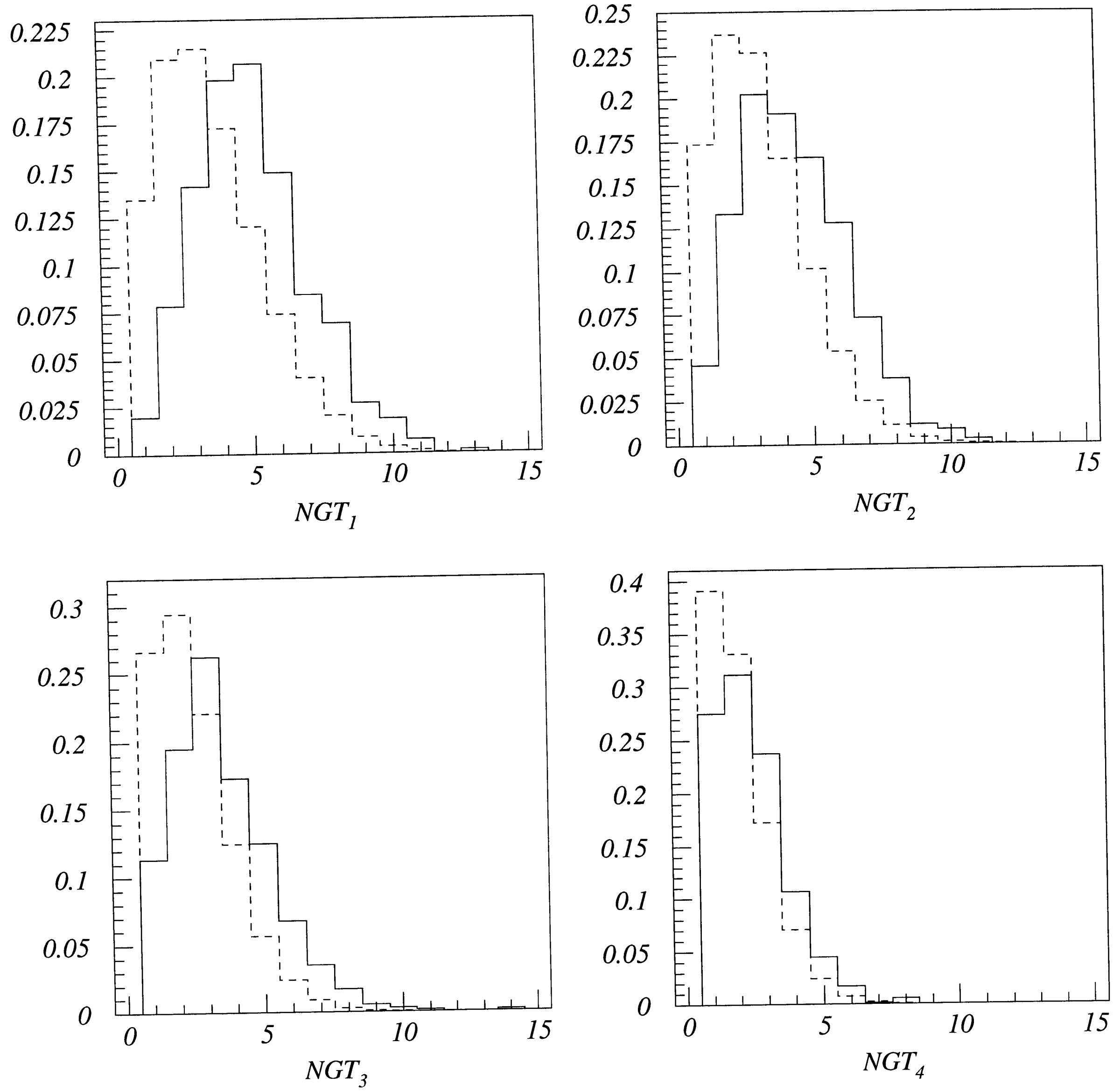


Figure 7: The normalized distributions of the number of tracks used to calculate the jet probabilities for the 4–b sample (solid line) and hadronic sample (dashed) after preselection.

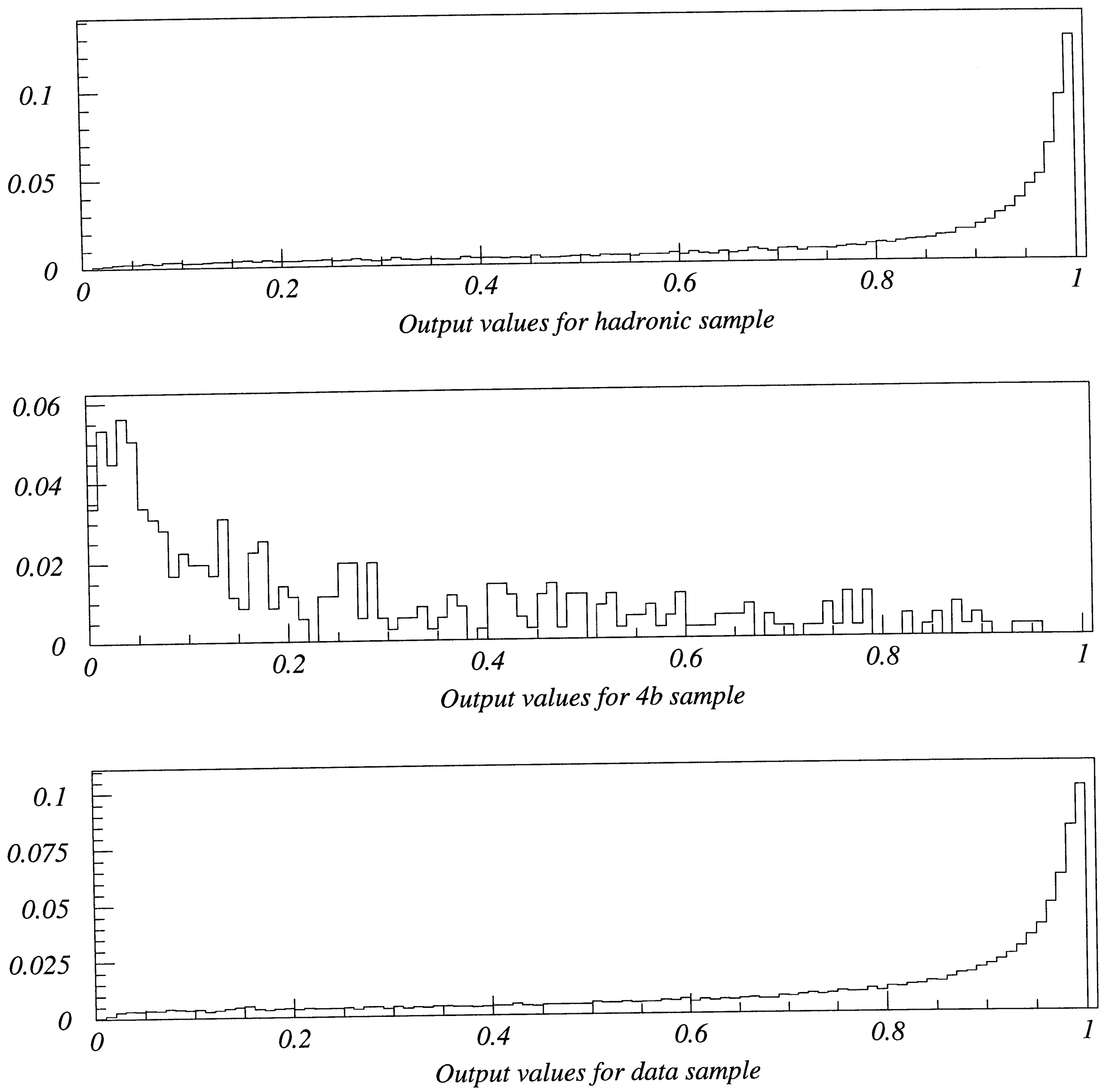


Figure 8: Typical distributions of the neural net output node values for the hadronic sample, the 4–b sample, and the data sample.

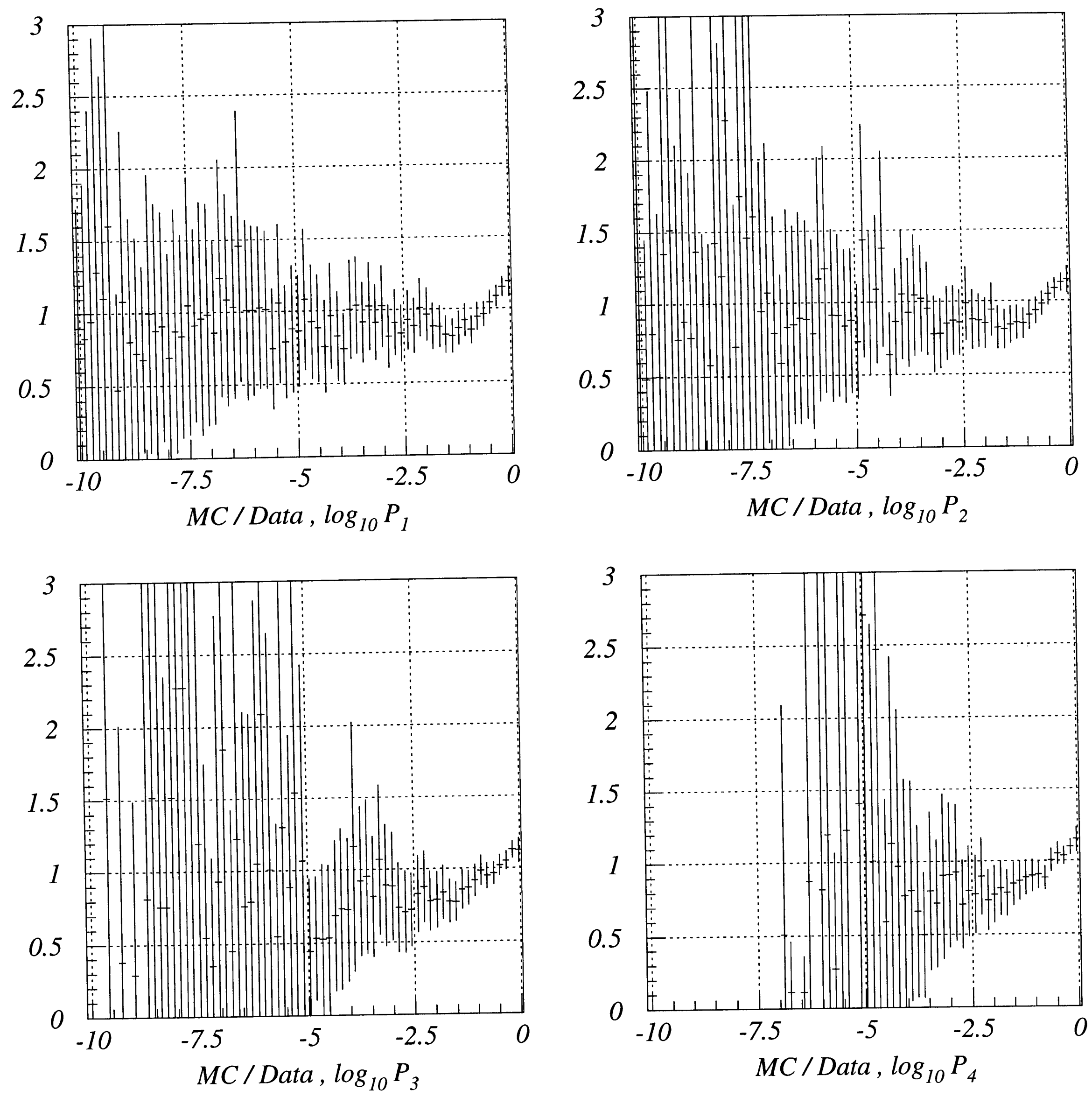


Figure 9: Ratio of normalized histograms in the four jetprobabilities (Monte Carlo divided by data).

References

[1] D. Brown, M. Frank, ALEPH note 92-135

[2] A. Ballestrero et al., Nucl. Phys. B415 (1994) 265.

[3] T. Sjöstrand, Computer Phys. Commun. 39 (1986) 347.

[4] C. Peterson et al., Computer Phys. Commun. 81 (1994) 185.