Looking for Instanton-induced Processes at HERA Using a Multivariate Technique Based on Range Searching

T. Carli[1] and B. Koblitz[1,2]

[1] *Universität Hamburg and* [2]*MPI für Physik, Föhringer Ring 6, D-80805 München*

Abstract. We present a method to discriminate instanton-induced processes from standard DIS background based on Range Searching. This method offers fast and automatic scanning of a large number of variables for a combination of variables giving high signal to background ratio and the smallest theoretical and experimental uncertainties.

INSTANTON-INDUCED PROCESSES

Instantons (1) are a fundamental non-perturbative aspect of QCD, inducing hard processes that are absent in perturbation theory. The expected cross section as calculated in "instanton-perturbation-theory" is sufficiently large (2, 3, 4) to make an experimental discovery possible (5, 6). For a more detailed introduction to instantons (I) see e.g. (7).

We study the prospect of a search for I-induced events modelled by the Monte Carlo Generator QCDINS (8) which generates I-induced events in deep-inelastic ep-scattering where a quark emerging from a $q\bar{q}$-splitting of the exchanged photon fuses with a gluon emitted from the proton. In the I-induced process $q\bar{q}$-pairs of each of the three light quark flavours and on average 2-3 gluons are produced. In the hadronic CMS they form a band (of about two units in pseudo-rapidity) of particles with high transverse energy which are homogeneously distributed in azimuth. Since in every event a pair of strange quarks is produced, in this band an increased number of kaons compared to standard DIS events is expected. Finally, the quark out of the split photon not participating in the instanton subprocess forms a hard jet.

The predicted cross section $\sigma^{(I)}_{\text{HERA}} = 29.2^{+9.9}_{-8.1}\,\text{pb}$ (4, 7) in a kinematic region where "instanton-perturbation-theory" ($x_B > 10^{-3}$, $0.1 < y < 0.9$, $Q^2 > 113\,\text{GeV}^2$) is applicable, is two orders of magnitude smaller than the DIS cross section $\sigma_{\text{DIS}} \approx 3000\,\text{pb}$. Therefore the highest possible signal to background ratio has to be achieved by exploiting observables characterising I-induced processes. To find these observables a large number of promising event variables have to be investigated and the sensitivity to systematic details in the modelling of the hadronic final state has to be tested. This requires a sophisticated and fast discrimination method to find the appropriate combination of event variables.

RANGE SEARCHING

Events can be classified as signal or background by estimating the probability density ρ of both these classes at the point of the event in the event-variable phase space, employing a Monte Carlo (MC) generator to sample the densities. In the case of neural networks (NN) this is done by fitting the probability-density with the adjusted weights of the neurons. To circumvent this time consuming procedure the density at each point can be directly estimated by counting the number of background and signal events in a surrounding box V. Given the ratio

$$\ell := \frac{\rho(I)}{\rho(DIS)} = \frac{\#I(V)}{\#DIS(V)}$$

the probability of an event to be a signal event is $D = \ell/(1+\ell)$. Compared to NN's this method also has the advantage of not extrapolating into phase space regions where there are no sample events available. Thus signals from data events outside the region covered by the MC simulation can be avoided. This is not the case for NN's which extrapolate into regions where there is no test data. Counting the number of events in the vicinity of a certain point is a problem known as Range Searching.

Range Searching algorithms have been developed which allow a search time $\sim \log(n)$, where n is the number of points that have to be searched (9). We employ an algorithm (11) suitable especially for a large number of events and dimensions (i.e. observables). The MC events are successively filled into the nodes of two binary trees, one for the signal events and one for the background events, where the criterion by which the decision is taken

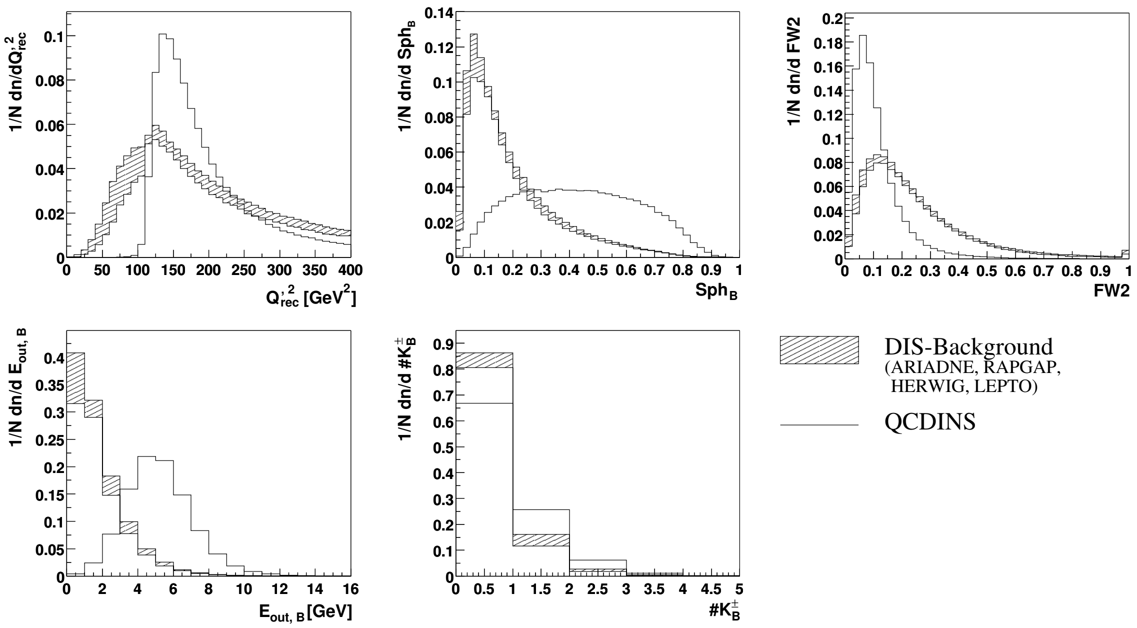

FIGURE 1. The characteristic event variables providing good instanton separation with small systematic uncertainties. Shown is the reconstructed virtuality of the quark entering the *I*-subprocess $Q'^2_{\rm rec}$, the sphericity of the particles in the *I*-Band in their rest system, the second Fox-Wolfram moment of these particles and the event shape variable $E_{\rm out,B}$ which is the projection of the particle transverse energy onto the axis, that makes this quantity maximal (see (6)). Finally the number of charged kaons in the *I*-Band is shown.

to descend to the left or right of a node is given by the value of one of the event variables. While descending the tree this variable cycles through all the ones considered. After filling, the position of every event in the tree is given by its coordinates in the event variable space. Classification of an event is done by searching in the trees for all background and signal events in the box V. This is done in the same manner as filling the tree.

RESULTS

Starting with 35 variables based on the hadronic final state the best 12 were chosen by calculating the discriminant with all 2-combinations (pairs) of the initial variables and taking those variables which provide a high separation power $S = \varepsilon(I)/\varepsilon(DIS)$ demanding an efficiency for instantons of $\varepsilon(I) = 10\%$. The number of considered variables is further reduced by calculating all 5-combinations and selecting those with highest separation power and a small systematic variation of the background. The systematic uncertainty was obtained by using four standard DIS-MC simulators which were tuned to data on representative hadronic final state quantities, in the range $Q^2 > 100\,\mathrm{GeV}^2$ at HERA (10). The variables forming the best combination is shown in Figure 1. The separation power for $\varepsilon(\mathrm{I}) = 10\%$ is $S = 126$. In Figure 2 the shape normalised discriminant D is shown for the *I*-induced and the background events, as well as the distribution for $D > 0.9$ normalised to a luminosity of $100\,\mathrm{pb}^{-1}$ which is comparable to that already collected by each of the HERA experiments H1 and ZEUS. An event sample can be isolated where half of the events are instantons while the *I*-efficiency is still 10%.

For a search method to be reliable and easy to apply it is important to have as few free parameters as possible. In the case of Range Searching these are the number of events in the search trees, the size of the neighbourhood V and the minimum number of events in this neighbourhood to classify an event. To reduce the number of parameters for the box size the ratios of the box edge lengths were fixed by defining a box which contains most of the events and letting V be a scaled version of this large box. The projections onto these box edges are shown in Figure 1. The variation of the result depending on the size of V is shown in Figure 3. Clearly the separation increases for smaller boxes with the number of events that populate the search trees, while for larger boxes this difference vanishes. The plateau is increasing in width with the number of tree events and reaches nearly an order of magnitude in size, thus allowing to use only an approximate size parameter and reducing the need for fine tuning, if enough MC statistics is available.

In addition a comparison with a single hidden layer feed forward NN was done. The network performing

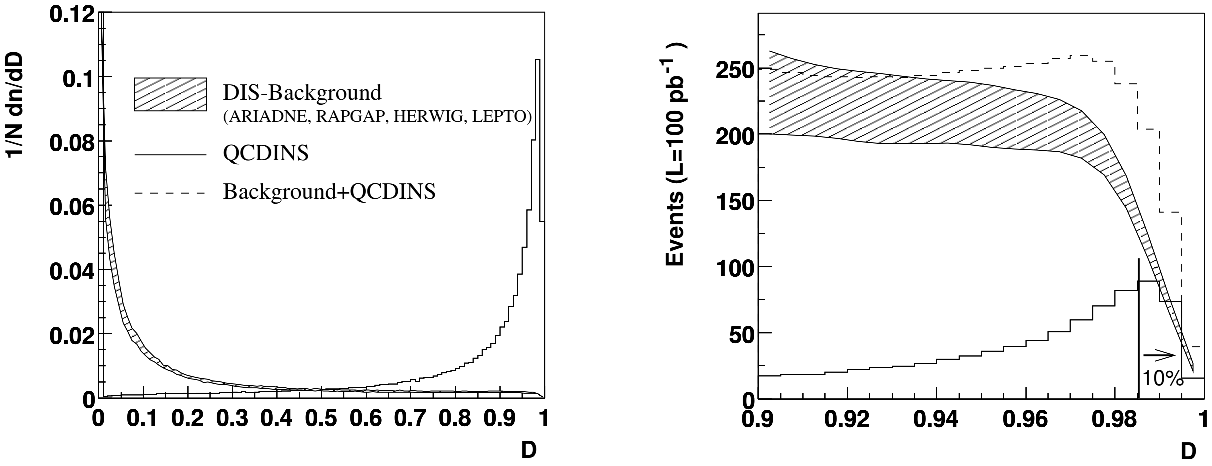

FIGURE 2. To the left, the shape normalised discriminant D for the instanton events using QCDINS and for standard DIS events using four MC simulators is shown. The second plot shows a zoom into the rightmost part and is now normalised to a luminosity of $L = 100\,\text{pb}^{-1}$. At $\varepsilon(I) = 10\%$ 178 I-induced events are expected.

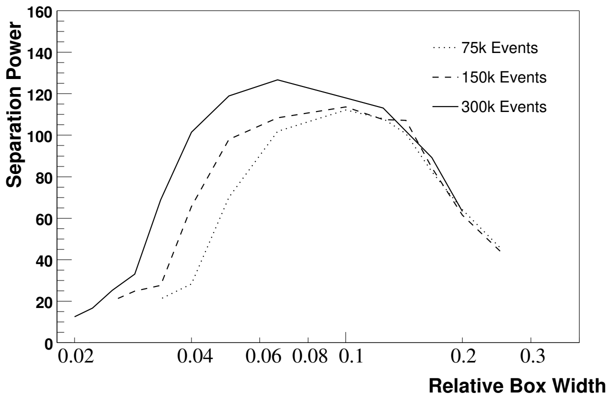

FIGURE 3. The separation power S at $\varepsilon(I) = 10\%$ for different box sizes and different numbers of events in the search trees.

best had 100 hidden nodes and was trained with the same input data. It reached a separation of $S = 116$ at an I-efficiency of 10%, being slightly worse than the Range Search method. Training the net was rather time consuming [1] and a lot of human intervention had to be done to adjust the input scales and training parameters.

CONCLUSIONS

The multivariate discrimination method based on Range Searching performs at least as good as a NN's when applied to the search for instantons at HERA. It is much less time consuming and can be easily used to automatically scan a large number of appropriate variables. The short processing time allows extensive searches for the best discriminating variables taking systematic effects into account. In a region where I-perturbation theory can be safely employed this novel discrimination method results in an 50% I-enriched data sample while the I-efficiency is still 10%.

[1] 4 h compared to 20 min for the Range Search method on a Linux PC

REFERENCES

1. A. Belavin et al., *Phys. Lett.* B **59**, 85 (1975); G. 't Hooft, *Phys. Rev.* D **14**, 3432 (1976).
2. S. Moch, A. Ringwald and F. Schrempp, *Nucl. Phys.* B **507**, 134 (1997).
3. A. Ringwald and F. Schrempp, *Phys. Lett.* B**438**, 217 ('98).
4. A. Ringwald and F. Schrempp, *Phys. Lett.* B**459**, 249 ('99).
5. A. Ringwald and F. Schrempp, hep-ph/9411217, in *Quarks '94*, eds. D. Yu. Grigoriev *et al.* (World Scientific, 1995).
6. T. Carli, J. Gerigk, A. Ringwald and F. Schrempp, hep-ph/9906441, in *Monte Carlo Generators for HERA Physics*, eds. A. T. Doyle *et al.*, DESY-PROC-1999-02.
7. A. Ringwald and F. Schrempp, hep-ph/9909338, in *New Trends in HERA Physics 1999*, eds. G. Grindhammer *et al.* (Springer, 2000).
8. A. Ringwald and F. Schrempp, hep-ph/9911516, in *Comput. Phys. Commun.* 132, 267 (2000)
9. R. Sedgewick, *Algorithms in C++*, Addison Wesley, 1992, Chapter 26
10. N. Brook *et al.*, in *Monte Carlo Generators for HERA Physics*, eds. A. T. Doyle *et al.*, DESY-PROC-1999-02.
11. Source available at `http://www.desy.de/~koblitz`