June 7, 2001

QUAERO

Manual

Bruce Knuteson
Lawrence Berkeley National Laboratory

Abstract

The use of QUAERO, a web-based tool that enables automatic searches for specific new phenomena, is described.

1 Introduction

It is generally recognized that the standard model, a successful description of the fundamental particles and their interactions, must be incomplete. Models that extend the standard model often predict rich phenomenology at the scale of a few hundred GeV, an energy regime accessible to the Fermilab Tevatron. Due in part to the complexity of the apparatus required to test models at such large energies, experimental responses to these ideas have not kept pace. Any technique that reduces the time required to test a particular candidate theory would allow more such theories to be tested, reducing the possibility that the data contain overlooked evidence for new physics.

Once data are collected and the backgrounds have been understood, the testing of any specific model in principle follows a well-defined procedure. In practice, this process has been far from automatic. Even when the basic selection criteria and background estimates are taken from a previous analysis, the reinterpretation of the data in the context of a new model often requires a substantial length of time.

Ideally, the data should be "published" in such a way that others in the community can easily use those data to test a variety of models. The publishing of experimental distributions in journals allows this to occur at some level, but an effective publishing of a multidimensional data set has, to our knowledge, not yet been accomplished by a large particle physics experiment. The problem appears to be that such data are very context-specific, requiring detailed knowledge of the complexities of the apparatus. This knowledge must somehow

be incorporated either into the data or into whatever tool the non-expert would use to analyze those data.

Many data samples and backgrounds have been defined in the context of SLEUTH [1], a quasi-model-independent search strategy for new high p_T physics that has been applied to a number of exclusive final states [2, 3] in the data collected by the DØ detector [4] during 1992–1996 in Run I of the Fermilab Tevatron. Using these samples and standard model backgrounds, this manual describes a tool (QUAERO) [5] that automatically optimizes an analysis for a particular signature.

2 Web form

The web form (`http://quaero.fnal.gov/`) has the following entries:

Final State. The appropriate final state is selected by choosing among those in the list. These final states are inclusive in the number of jets: $e\mu(\not{E}_T)(nj)$ indicates zero or more jets, and $e\not{E}_T\,2j(nj)$ and $ee\,2j(nj)$ indicate two or more jets.

Smear? Check this box if your signal file has not been smeared for DØ detector resolution. The energy resolution of electromagnetic objects (electrons and photons) in Run I is $\delta E/E = 15\%/\sqrt{E} \oplus 0.3\%$; the energy resolution of hadronic objects (jets) is $\delta E/E = 80\%/\sqrt{E}$; and the momentum resolution of muons is $\delta(1/p) = 0.18(p-2)/p^2 \oplus 0.003$. Here E and p are in units of GeV, and $\oplus$ means addition in quadrature.

Pythia input The signal process can be specified either by providing commands that will allow PYTHIA [6] to generate this process, or by providing a text file with the 4-vectors of all objects in a number of simulated events (discussed below). If generation of the signal within PYTHIA is desired, the radio button in front of "Pythia input" should be selected. PYTHIA may be used to model initial and final state radiation and quark fragmentation, in addition to generating parton-level events. Pythia v6.156 is used within Quaero. Ten thousand Monte Carlo events are generated by default; a similar number of events should be provided if a signal file is used instead.

The response of the DØ calorimeter to the collimated products of fragmentation (jets) is simulated by clustering final state particles. This approach has been shown (through studies within DØ and in the course of the Run II SUSY/Higgs Workshop) to give results comparable to those obtained with a complete (GEANT-based) simulation of the DØ detector. The algorithm used for clustering finds all jets with $|\eta| < 3$ and $p_T > 10$ GeV. Final state particles are clustered using a simple cone algorithm with cone radius of 0.7 in η–ϕ; no

further splitting or merging of jets is attempted. Fragmentation of final state particles is turned off by default.

The PYTHIA commands currently recognized are `msub`, `pmas`, `ckin`, `brat`, `mdme`, `imss`, and `rmss`. The meanings of these parameters are detailed in documentation available from the Pythia home page [6]. For example, to specify standard model Higgs production with an assumed Higgs boson mass of 200 GeV, we set `msub(102)=1` and `pmas(25,1)=200`.

Signal File. QUAERO will also accept as input a text file with the 4-vectors of all objects in a number of simulated signal events. If this option is desired, the radio button in front of "signal file" should be selected. Each row of this file represents one Monte Carlo signal event, and consists of several strings and numbers separated by spaces or tabs. Each row must begin with a number representing that event's weight relative to the rest of the events in the file. After the event's weight, each object in the event should then be listed by providing the type of object (**e** for electron, **mu** for muon, **j** for jet, **met** for missing p_T, and **uncl** for unclustered energy), followed by the x, y, and z components of the momentum of the object. Momenta should be given in standard C or FORTRAN notation (e.g., 107.8 or 0.1078e+3) in units of GeV. `met_pz` and `uncl_pz` are not measured in a $p\bar{p}$ collider, and should be set equal to zero. In order to ensure that all objects in each event have been correctly accounted for, QUAERO checks for transverse momentum balance in each event. The end of an event is signaled by the presence of a semicolon, separated by whitespace. A row in a signal file for the final state $e\mu(\not{E}_T)(nj)$ might therefore look like this:
`weight` **e** `e_px e_py e_pz` **mu** `mu_px mu_py mu_pz` **j** `j1_px j1_py j1_pz` **j** `j2_px j2_py j2_pz` **met** `met_px met_py 0.` **uncl** `uncl_px uncl_py 0.` **;**

xsec. If "signal file" is selected, the cross section corresponding to the input file should be entered in this box (in units of picobarns). If "Pythia input" is selected, the production cross section is determined by Pythia and this box is ignored.

View. If "View" is selected, the result of QUAERO's analysis will be one-dimensional plots of signal, background, and data in the specified variables.

Search. If "Search" is selected, the result of QUAERO's analysis will be cross section limits for the signal at 50%, 90%, and 95% confidence levels, in addition to one-dimensional plots of signal, background, and data in the specified variables.

Backgrounds. Each checked process will be included in the background estimate. One or more processes may be excluded from the background estimate

by unchecking the appropriate box(es). If "Search" is selected, the signal will be searched for over a background estimate including only those processes checked.

Constraints. Specific cuts may be inserted here in standard C++ notation. The cut `mass(e,met)>110` could be imposed to remove W-related backgrounds, for example. In general, we advocate not imposing any specific cuts, and leaving this box blank. In this example, including `mass(e,met)` as a variable instead of using it as a constraint allows QUAERO more flexibility in determining an optimal region.

Variables. An object is denoted by `o#`, where `o` = `e`, `mu`, `j`, or `met` (but not `uncl`), and `#` is a number representing its ordering in transverse momenta. Thus `e1` denotes the electron with largest transverse momentum, and `e2` the electron with second largest transverse momentum, and so on. If there is only one object of type `o`, `o` can be used in place of `o1`, and `met` is used rather than `met1`.

Kinematic information of each object `o` can be accessed using `o_`κ, where κ can be any of `px`, `py`, `pt`, `pz`, `e`, `phi`, `eta`, or `theta`. Thus `e1_pt` is the transverse momentum of the leading electron, and `j2_eta` is the pseudorapidity of the second leading jet. Objects can be combined by addition: the p_T of a W boson can be obtained by combining the electron and missing transverse energy into one object and then taking accessing its transverse momentum with the syntax `(e+met)_pt`, for example.

QUAERO is able to parse most standard C notation and functions. Thus `e1_pt+mu1_pt`, `exp(-abs(j1_eta-j2_eta))`, and `abs(e_phi-mu_phi)>pi?2*pi-abs(e_phi-mu_phi):abs(e_phi-mu_phi)` are all legal variable names. Other commands that QUAERO understands include:

- `deltaphi(o#,o#)` returns the azimuthal distance between two objects.
- `mass(o#,...)` returns the invariant mass of the `o#`. Thus `mass(e1,e2,j1,j2)` computes the invariant mass of two electrons and two jets.
- `transversemass(o#,...)` computes the transverse mass of its arguments.
- `aplanarity()` computes the aplanarity of the event.
- `ChiSqdConstrain(o#,o#',V)` fits the two objects `o#` and `o#'` as the decay of the particle `V`, where `V` = `W` or `Z`. `ChiSqdConstrain(o#,o#',m,w)` fits the two objects `o#` and `o#'` as the decay of a particle with mass `m` and width `w`. The functions return the χ^2 of the fit. If either of the objects is `met`, this constraint determines `met_pz`. If this function is used in an entry to "Constraints" or to the first variable ("v1"), the fitted (rather than the original) objects are used when computing the remaining variables.

Each of these commands can be used together with other expressions and functions. Examples of use in the context of actual searches are available from the web page.

The choice of variables to use will in many cases be apparent from the characteristics of the signal. In the absence of such "obvious" variables, we suggest the variable sets specified in Ref. [1]. The legality of each provided variable is tested by QUAERO before the request is submitted; invalid variables prompt a friendly error message. Up to three variables may be used.

Requestor Name, Institution, and Email. The results of the query will be returned by email to the address provided here.

Submit. Clicking this button submits the request.

3 Algorithm

The signals predicted by most theories of physics beyond the standard model involve an increased number of predicted events in some region of an appropriate variable space. In this case the optimization of the analysis can be understood as the selection of the region in this variable space that minimizes $\overline{\sigma^{95\%}}$, the expected 95% confidence level (CL) upper limit on the cross section of the signal in question, assuming the data contain no signal. The optimization algorithm consists of a few simple steps:

(i) Kernel density estimation [7] is used to estimate the probability distributions $p(\vec{x}|s)$ and $p(\vec{x}|b)$ for the signal and background samples in a low-dimensional variable space $\mathcal{V}$, where $\vec{x} \in \mathcal{V}$. The signal sample is contained in a Monte Carlo file provided as input to QUAERO. The background sample is constructed from all known standard model and instrumental sources.

(ii) A discriminant function $D(\vec{x})$ is defined by [8]

$$D(\vec{x}) = \frac{p(\vec{x}|s)}{p(\vec{x}|s) + p(\vec{x}|b)}. \tag{1}$$

The semi-positive-definiteness of $p(\vec{x}|s)$ and $p(\vec{x}|b)$ restricts $D(\vec{x})$ to the interval $[0, 1]$ for all $\vec{x}$.

(iii) The *sensitivity* $\mathcal{S}$ of a particular threshold D_{cut} on the discriminant function is defined as the reciprocal of $\overline{\sigma^{95\%}}$. D_{cut} is chosen to maximize $\mathcal{S}$.

(iv) The region of variable space having $D(\vec{x}) > D_{\text{cut}}$ is used to determine the actual 95% CL cross section upper limit $\sigma^{95\%}$.

4 Results

The result of this analysis is then emailed to the address provided in the request. This email contains:

(i) the sensitivity $\mathcal{S}$ of the analysis, in units of pb^{-1};
(ii) the number of signal and background events expected in the selected region, and the number of data events observed; and
(iii) cross section $\times$ branching fraction limits at the 50%, 90%, and 95% confidence level.

The QUAERO reference that should be cited when publishing the result is also included.

If a signal is analyzed using several variable combinations, the result corresponding to the largest sensitivity should be considered to be the final answer. This practice eliminates any bias introduced when selecting among which of several analyses to publish.

If a non-negligible excess is observed, a cross section central value and upper and lower errors will be returned in addition to limits at the 90% and 95% confidence level.

5 Conclusions

We have presented QUAERO, a web-based tool that enables automatic searches for specific new phenomena. A number of examples are available for experimentation from the Quaero web site. Enjoy!

Acknowledgments

Daniel Whiteson assisted in the debugging of the QUAERO code and interface. Greg Landsberg contributed several examples.

References

[1] DØ Collaboration, B. Abbott *et al.*, Phys. Rev. D **62**, 92004 (2000).

[2] DØ Collaboration, B. Abbott *et al.*, Phys. Rev. Lett., **86**, 3712 (2001).

[3] DØ Collaboration, B. Abbott *et al.*, Phys. Rev. D **64**, 012004 (2001).

[4] DØ Collaboration, S. Abachi *et al.*, Nucl. Instr. and Methods in Phys. Res. A **338**, 185 (1994).

[5] DØ Collaboration, V. Abazov *et al.*, submitted to Phys. Rev. Lett., hep-ex/0106039 (2001).

[6] T. Sjöstrand, Comput. Phys. Commun. **82**, 74 (1994); `http://www.thep.lu.se/~torbjorn/Pythia.html`.

[7] D. Scott, *Multivariate Density Estimation*, John Wiley & Sons, 1992.

[8] L. Holmström, S. R. Sain, and H. E. Miettinen, *Computer Physics Communications*, 88:195–210, 1995.