The QUAERO Algorithm

Bruce Knuteson*
Massachusetts Institute of Technology
(Dated: December 12, 2005)

This article describes the algorithm used by QUAERO to automate the testing of specific hypotheses against high-p_T data.

Contents

I. INTRODUCTION

An algorithm for testing an arbitrary hypothesis against high-p_T data exists — it is implemented by experimental physicists in meetings, in hallway discussions, and at their terminals. It is natural to wonder whether gains in efficiency and robustness might be achieved by streamlining this algorithm, making it completely prescriptive, and implementing it in code.

Ref. [1] specifies the QUAERO interface in detail. The interface has a back end to collider experiments, into which each collaboration inserts its data and expert knowledge. Each participating collaboration provides

- data, in the form of the four-vectors of all high-p_T objects seen in all events collected in the detector;
- backgrounds, in the form of the four-vectors of all high-p_T objects seen in Monte Carlo events predicted by the Standard Model;
- events that have been run through the experiment's detector simulation, so that an algorithm called TURBOSIM [2] can learn the detector response;
- systematic errors, including specification of the sources, the correlations among those sources, and their effect on each four-vector quantity; and
- specification of any requisite post-processing.

A physicist presents a hypothesis to QUAERO's front end. The hypothesis is some "signal," in the form of commands to one of the standard event generators. Interfaces to PYTHIA [3], SUSPECT [4], and MADEVENT [5] are currently supported. This signal, together with whatever background processes the physicist wishes to include, defines the hypothesis to be tested. QUAERO's response is a single number quantifying the extent to which the data (dis)favor that hypothesis, relative to the Standard Model.

Section II describes the steps in the QUAERO algorithm, using a measurement of the Z boson mass in 1 pb^{-1} of Tevatron Run II data as an illustrative example. Comparison of QUAERO's performance with traditional means requires the performance of parallel analyses on real data, handling all details; this work is in progress, and the subject of articles in preparation. The present article provides a clean introduction to the inner workings of QUAERO in a toy scenario, allowing focus on the algorithm itself. In the examples provided, all Standard Model processes have been included in the background estimate, with realistic object identification efficiencies, geometric acceptances, and object misidentification probabilities. To keep this article standalone, the Pretty Good Simulation (PGS) [6] provides a rough parametrized detector simulation appropriate for either of the two Tevatron experiments, and 1 pb^{-1} of "data" is drawn from the Standard Model expectation, rather than from physical collisions.

II. ALGORITHM

For a particular hypothesis $\mathcal{H}$, the quantity of interest is $\log_{10}\mathcal{L}(\mathcal{H})$, where

$$\mathcal{L}(\mathcal{H}) = \frac{p(\mathcal{D}|\mathcal{H})}{p(\mathcal{D}|\mathrm{SM})}, \qquad (1)$$

*URL: http://mit.fnal.gov/~knuteson/; Electronic address: knuteson@mit.edu

$\mathcal{D}$ are the data, and SM is the Standard Model.

The computation of this quantity requires

- devising an analysis strategy that respects the allotted time,
- generating the events predicted by $\mathcal{H}$,
- simulating the response of each detector to these events,
- partitioning the events into exclusive final states,
- choosing variables for each final state in each experiment,
- binning each variable space,
- calculating a binned likelihood for each final state in each experiment,
- combining these likelihoods for all final states within an experiment,
- combining these likelihoods among experiments, and
- incorporating systematic errors.

This section considers each of these steps in turn, concluding with a brief discussion on the interpretation of results. The measurement of M_Z in 1 pb^{-1} of Tevatron Run II data is used as a toy example for illustrative purposes throughout.

A. Time

The querying physicist provides a target time for analysis completion, in units of kiloseconds. QUAERO is designed to adjust its analysis strategy to perform the most sensitive achievable test of the provided hypothesis within the allotted time.

QUAERO can better learn the shapes of distributions by generating more signal events; it can construct better kernel estimates from these signal events if allowed more attempts; it can make use of a larger subset of the data if allowed the time to access more events; and it can perform a more robust integration over systematic errors if allowed time to sample more points in the space of systematic shifts. The time cost of QUAERO's analysis as a function of these four parameters (the number of generated signal events, the number of starting points for the construction of kernel estimates, the number of individual Standard Model background events that QUAERO is allowed to touch, and the number of points sampled in the space of systematic shifts) has been emperically determined. QUAERO begins its analysis by optimizing a figure of merit as a function of these four parameters, respecting the provided time constraint.

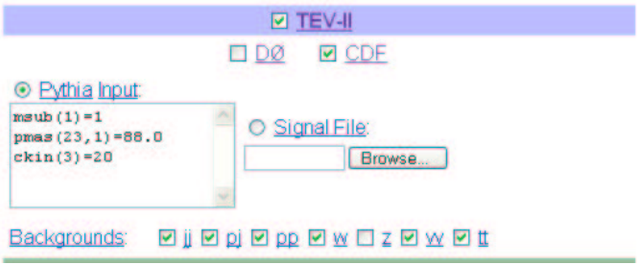

FIG. 1: An example of PYTHIA input to QUAERO. The first line selects Z/γ^* production; the second line sets the Z boson mass to 88 GeV; the third line restricts the generation to hard interactions. These events replace previously generated Z events with $M_Z = 91.2$ GeV, removed from the background estimate by the unchecked box.

B. Event generation

If the signal is provided in the form of commands to PYTHIA, QUAERO uses PYTHIA to generate the predicted events. The number of events generated corresponds to an integrated luminosity one hundred times that collected. To more clearly illustrate the effect of limited Monte Carlo statistics, an integrated luminosity ten times that collected will be used for the example in this section.

An example of PYTHIA input to QUAERO is provided in Fig. 1; this section describes in detail QUAERO's analysis of this input. QUAERO passes these commands directly to PYTHIA, which creates a STDHEP file with generated events. These events, together with events from all included background processes, define the hypothesis $\mathcal{H}$ to be tested. Previously generated events from all Standard Model processes, provided by each experiment, define the reference model SM to which the hypothesis will be compared.

C. Detector simulation

A TURBOSIM detector simulation takes the STDHEP file with generated events, simulates the response of the detector to those events, and produces a file in QUAERO format with the four-vectors of reconstructed objects.

D. Final states

Most collider analyses are performed on inclusive final states. QUAERO takes the more natural view that events containing different objects ($e^{\pm}$, $\mu^{\pm}$, $\tau^{\pm}$, γ, j, b, $\not{p}$) are fundamentally dissimilar, and should be treated separately.

The events predicted by the hypothesis $\mathcal{H}$ are partitioned into exclusive final states according to the objects observed in each event. The events predicted by the ref-

erence model SM and the events observed in the data $\mathcal{D}$ are similarly partitioned. This partitioning is orthogonal; each event is placed in one and only one exclusive final state.

QUAERO determines which final states to consider by ordering the final states according to decreasing $s/(\sqrt{b}\,N_{MC})$, where s is the sum of the weights of the signal events in a particular final state, b is the sum of the weights of the Standard Model events in that final state, and N_{MC} is the number of Standard Model monte carlo events QUAERO would need to touch if it decides to consider that final state. Starting from the top of the list, QUAERO adds each final state to the set of final states it will consider if the total number of events it will be considering is smaller than the total number of events QUAERO decided it has time to consider in Sec. II A. In the example of this section, QUAERO realizes the final states e^+e^- and $\mu^+\mu^-$ (among others) need to be considered.

E. Variables

The partitioning of the data into exclusive final states enables a straightforward method of variable selection.

Hadron collider events with n final state objects populate a $3n-2$ dimensional space, where typically $2 \leq n \lesssim 6$. The full-dimensional space usually cannot be reliably modeled with the limited number N_{MC} of Monte Carlo events at hand. Attention is thus restricted to a d-dimensional subspace, where $d = \lfloor \log_{10} N_{MC} \rfloor$. [9]

For a hadron collider experiment, the variables considered in each final state are

- the transverse momentum (p_T) of each object,
- the pseudorapidity (η) of each object,
- the distance in azimuthal angle ($\Delta\phi$) between each object pair,
- the distance ($\Delta\mathcal{R}$) in pseudorapidity and azimuth of each object pair, and
- the invariant masses of all combinations of two or more objects.

For lepton collider experiments the same list holds, with the substitution of energy (E) for transverse momentum, and polar angle (θ) for pseudorapidity.

These variables are ordered according to decreasing values of the Kolmogorov-Smirnov (KS) statistic

$$\max_{x_0} \left| \int_{-\infty}^{x_0} p(x|\mathcal{H}) - \int_{-\infty}^{x_0} p(x|\mathrm{SM}) \right|, \qquad (2)$$

the maximal difference between the cumulative distribution functions of $\mathcal{H}$ and SM in the variable x. Beginning with the first variable in this ordering and continuing until d variables have been chosen, each variable is added to those considered unless the smallest eigenvalue of the correlation matrix of this variable and the $q-1$ variables already chosen is smaller than $1/q$.

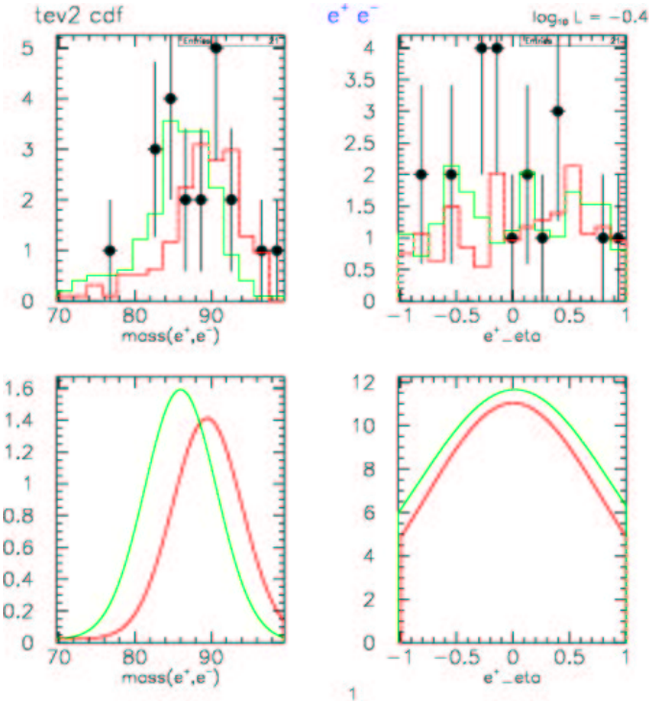

FIG. 2: (Upper) Histograms of the prediction of the hypothesis $\mathcal{H}$ (green) and of the Standard Model SM (red), together with the data $\mathcal{D}$ (solid points), in QUAERO's chosen variables in the final state e^+e^-. QUAERO's variable selection algorithm picks the invariant mass of the two electrons `mass(e+,e-)` (left) as the first variable, and selects the pseudorapidity of the positron (right) as an inferior second variable. The vertical axis shows the number of events predicted in each bin. (Lower) One-dimensional projections of the kernel estimates, obtained by integrating the two-dimensional densities in Fig. 3 over the other variable. The vertical axis shows the number of predicted events per GeV (left) and per unit of pseudorapidity (right). In all four panes, the integral of the red curve is 18 events, the number predicted in this final state by SM. The integral of the green curve in all four panes is 20 events, the number predicted in this final state by $\mathcal{H}$. As expected, the Z boson cross section $\sigma(p\bar{p} \to Z)$ is larger for smaller values of M_Z.

In the example at hand, the final state e^+e^- contains $\approx$ 200 SM Monte Carlo points, allowing the use of two variables. QUAERO chooses $m_{e^+e^-}$ as the first variable, as expected, and picks the positron's pseudorapidity as a second variable. Figure 2 shows the prediction of the hypothesis $\mathcal{H}$, the reference model SM, and the data $\mathcal{D}$ in these variables. [10] In the final state $\mu^+\mu^-$, QUAERO chooses $m_{\mu^+\mu^-}$ as the first variable, and the azimuthal angle of the positively charged muon as a throw-away second variable.

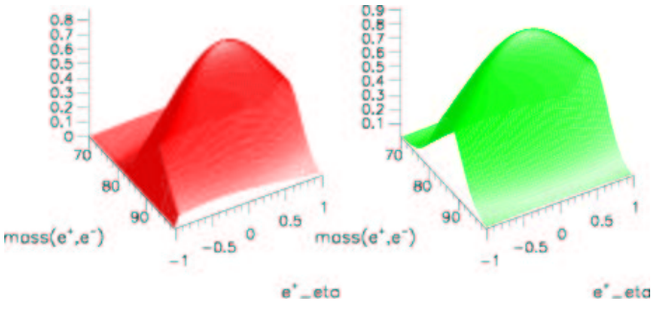

FIG. 3: Two-dimensional kernel estimates in QUAERO's chosen variable space for the prediction of the reference model SM (red, left) and of the hypothesis $\mathcal{H}$ (green, right) in the final state e^+e^-. The estimates are intrinsically multivariate, and are not simply products of one-dimensional densities, although the difference is moot for loosely correlated variables such as these.

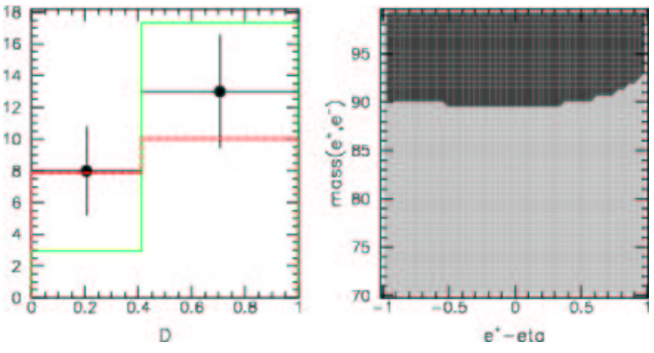

FIG. 4: (Left) The optimally-binned histogram of the discriminant D for the predictions of SM (red), $\mathcal{H}$ (green), and for the data $\mathcal{D}$ (solid points). (Right) The mapping of the bins in D back into regions in the original variable space. The dark region corresponds to points $\vec{x}$ in the variable space for which $D(\vec{x}) < 0.4$; the light region corresponds to points $\vec{x}$ in the variable space for which $D(\vec{x}) > 0.4$. Recall that $\mathcal{H}$ differs from SM in asserting $M_Z = 88.0$ GeV; thus the light region, for which the discriminant is large, is roughly the region $m_{e^+e^-} < 90$ GeV.

F. Kernel estimate

The discrete Monte Carlo events predicted by $\mathcal{H}$ and SM in the chosen variable space in each final state are used to construct smooth estimates using FEWKDE, a fast kernel density estimation algorithm [7]. FEWKDE provides multivariate density estimates $p(\vec{x}|\mathcal{H})$ and $p(\vec{x}|\text{SM})$, where $\vec{x}$ denotes an arbitrary point in the chosen variable space. Figure 3 shows the two-dimensional kernel estimates for the predictions of $\mathcal{H}$ and SM in the variable space of $m_{e^+e^-}$ and positron pseudorapidity in the final state e^+e^-. The lower panels of Fig. 2 show the one-dimensional projections, obtained by integrating out the other variable, for comparison with the histograms in the upper panels of Fig. 2.

G. Binning

QUAERO then forms the discriminant

$$D(\vec{x}) = \frac{p(\vec{x}|\mathcal{H})}{p(\vec{x}|\mathcal{H}) + p(\vec{x}|\text{SM})} \tag{3}$$

from the two density estimates $p(\vec{x}|\mathcal{H})$ and $p(\vec{x}|\text{SM})$. The discriminant $D(\vec{x})$ takes on values between zero and unity, approaching zero in regions in which the number of events predicted by the reference model SM greatly exceeds the number of events predicted by the hypothesis $\mathcal{H}$, and approaching unity in regions in which the number of events predicted by $\mathcal{H}$ greatly exceeds the number of events predicted by SM.

The value of the discriminant D at the position of each Monte Carlo event predicted by $\mathcal{H}$ is computed, together with the value of D at each Monte Carlo event predicted by SM. The resulting distributions in D are binned using a prescription for optimal binning [8].

The resulting binned histogram in D in the example of this section in the final state e^+e^- is shown in Fig. 4. QUAERO chooses to consider only two bins, placing a bin edge at $D = 0.4$. As advertised, SM Monte Carlo events tend to lie at variable values $\vec{x}$ for which $D(\vec{x})$ is small, and $\mathcal{H}$ Monte Carlo points tend to lie at variable values $\vec{x}$ for which $D(\vec{x})$ is large. From the red histogram in the left panel of Fig. 4, the reference model SM predicts 18.0 events in the final state e^+e^-; of these, 8.0 events are predicted to lie in regions of the chosen variable space for which the discriminant D is less than 0.4, and 10 events are predicted to lie in regions for which the discriminant is greater than 0.4. From the green histogram in the same panel, the hypothesis $\mathcal{H}$ predicts 3.0 events in regions of variable space with $D < 0.4$, and 17.0 events in regions with $D > 0.4$. The discriminant evaluated at the positions of the events observed in the data yield 8 data events with $D < 0.4$, and 13 data events with $D > 0.4$.

The right panel of Fig. 4 shows how the two bins in the discriminant map back onto the original variable space of $m_{e^+e^-}$ and positron pseudorapidity. The dark region corresponds to points $\vec{x}$ in the variable space for which $D(\vec{x}) < 0.4$; the light region corresponds to points $\vec{x}$ in the variable space for which $D(\vec{x}) > 0.4$. In the reference model SM, $M_Z = 91.2$ GeV; in $\mathcal{H}$, $M_Z = 88.0$ GeV. The boundary between the light and dark regions, defined by those points $\vec{x}$ for which $D(\vec{x}) \approx 0.4$, is at roughly $m_{e^+e^-} = 90$ GeV, in accord with intuition. Corresponding figures for the final state $\mu^+\mu^-$ are similar.

In both the final states e^+e^- and $\mu^+\mu^-$ QUAERO uses only two bins in the discriminant because of limited Monte Carlo statistics, and because the predictions of $\mathcal{H}$ and SM are not widely different. Dozens of bins may be used if the predictions are sufficiently different to warrant the finer binning, and if QUAERO has enough Monte Carlo points to robustly estimate the predicted number of events from $\mathcal{H}$ and from SM in each bin. Nothing is signif-

icant about the placement of the bin edge at $D = 0.407$; QUAERO optimizes the position of the bin edges on its own.

H. Likelihood

The probability of observing the data $\mathcal{D}$ in a particular final state (fs) within a particular experiment (exp), assuming the correctness of the hypothesis $\mathcal{H}$ and a vector $\vec{s}$ of systematic offsets, is

$$p(\mathcal{D}_{(\mathrm{exp})(\mathrm{fs})}|\mathcal{H},\vec{s}) = \prod_{i=1}^{N_{\mathrm{bins}}} \frac{e^{-h_i} {h_i}^{N_i}}{N_i!}, \quad (4)$$

where h_i is the number of events predicted by $\mathcal{H}$ in the i^{th} bin in this final state in this experiment, and N_i is the number of data events observed in that bin. Similarly,

$$p(\mathcal{D}_{(\mathrm{exp})(\mathrm{fs})}|\mathrm{SM},\vec{s}) = \prod_{i=1}^{N_{\mathrm{bins}}} \frac{e^{-b_i} {b_i}^{N_i}}{N_i!}, \quad (5)$$

where b_i is the number of events predicted by SM in the i^{th} bin in this final state in this experiment.

In this example, ignoring the uncertainty in the prediction in each bin due to finite Monte Carlo statistics,

$$\begin{aligned} p(\mathcal{D}_{(\mathrm{CDF})(e^+e^-)}|\mathcal{H},\vec{s}) &= \frac{e^{-3.0}3.0^8}{8!} \times \frac{e^{-17.0}17.0^{13}}{13!} \quad (6)\\ &= 0.0081 \times 0.066 \\ &= 0.00053, \end{aligned}$$

and

$$\begin{aligned} p(\mathcal{D}_{(\mathrm{CDF})(e^+e^-)}|\mathrm{SM},\vec{s}) &= \frac{e^{-8.0}8.0^8}{8!} \times \frac{e^{-10.0}10.0^{13}}{13!} \quad (7)\\ &= 0.14 \times 0.073 \\ &= 0.010, \end{aligned}$$

indicating that the data $\mathcal{D}$ favor SM relative to $\mathcal{H}$ by a factor of $0.010/0.0005 = 20$ in the e^+e^- final state at CDF. After accounting for the significant uncertainty ($\approx \pm 5$ events) in the prediction in each bin due to limited Monte Carlo statistics, the evidence provided by $\mathcal{D}$ in support of SM relative to $\mathcal{H}$ is only a factor of 2.5. Here as elsewhere, QUAERO's performance improves with the number of Monte Carlo points at its disposal.

I. Combination of final states

Probabilities $p(\mathcal{D}_{(\mathrm{exp})(\mathrm{fs})}|\mathcal{H},\vec{s})$ from individual final states are combined into a probability $p(\mathcal{D}_{(\mathrm{exp})}|\mathcal{H},\vec{s})$ for the experiment by multiplication:

$$p(\mathcal{D}_{(\mathrm{exp})}|\mathcal{H},\vec{s}) = \prod_{\mathrm{fs}} p(\mathcal{D}_{(\mathrm{exp})(\mathrm{fs})}|\mathcal{H},\vec{s}). \quad (8)$$

Similarly,

$$p(\mathcal{D}_{(\mathrm{exp})}|\mathrm{SM},\vec{s}) = \prod_{\mathrm{fs}} p(\mathcal{D}_{(\mathrm{exp})(\mathrm{fs})}|\mathrm{SM},\vec{s}). \quad (9)$$

In this example,

$$\begin{aligned} p(\mathcal{D}_{(\mathrm{CDF})}|\mathcal{H},\vec{s}) = \; & p(\mathcal{D}_{(\mathrm{CDF})(e^+e^-)}|\mathcal{H},\vec{s}) \times \quad (10)\\ & p(\mathcal{D}_{(\mathrm{CDF})(\mu^+\mu^-)}|\mathcal{H},\vec{s}) \times \\ & (\text{other final states}) \end{aligned}$$

and similarly for $p(\mathcal{D}_{(\mathrm{CDF})}|\mathrm{SM},\vec{s})$. Factors in Eq. 10 from final states for which the predictions of $\mathcal{H}$ and SM are similar cancel when the ratio of likelihoods is taken.

J. Combination of experiments

Probabilities $p(\mathcal{D}_{(\mathrm{exp})}|\mathcal{H},\vec{s})$ from each experiment are combined into a total probability $p(\mathcal{D}|\mathcal{H},\vec{s})$ by another multiplication:

$$p(\mathcal{D}|\mathcal{H},\vec{s}) = \prod_{\mathrm{exp}} p(\mathcal{D}_{(\mathrm{exp})}|\mathcal{H},\vec{s}). \quad (11)$$

Similarly,

$$p(\mathcal{D}|\mathrm{SM},\vec{s}) = \prod_{\mathrm{exp}} p(\mathcal{D}_{(\mathrm{exp})}|\mathrm{SM},\vec{s}). \quad (12)$$

In this example, assuming data from both Tevatron experiments,

$$p(\mathcal{D}|\mathcal{H},\vec{s}) = p(\mathcal{D}_{(\mathrm{CDF})}|\mathcal{H},\vec{s}) \times p(\mathcal{D}_{(\mathrm{D\O})}|\mathcal{H},\vec{s}); \quad (13)$$

similarly,

$$p(\mathcal{D}|\mathrm{SM},\vec{s}) = p(\mathcal{D}_{(\mathrm{CDF})}|\mathrm{SM},\vec{s}) \times p(\mathcal{D}_{(\mathrm{D\O})}|\mathrm{SM},\vec{s}); \quad (14)$$

K. Systematic errors

Systematic errors are incorporated by repeating the above steps many times with different systematic offsets $\vec{s}$, allowing the Monte Carlo computation of the integrals

$$p(\mathcal{D}|\mathcal{H}) = \int p(\mathcal{D}|\mathcal{H},\vec{s}) p(\vec{s})\, d\vec{s} \quad (15)$$

and

$$p(\mathcal{D}|\mathrm{SM}) = \int p(\mathcal{D}|\mathrm{SM},\vec{s}) p(\vec{s})\, d\vec{s}. \quad (16)$$

The final number of interest is the ratio of these likelihoods,

$$\mathcal{L}(\mathcal{H}) = \frac{p(\mathcal{D}|\mathcal{H})}{p(\mathcal{D}|\mathrm{SM})}. \quad (17)$$

The data $\mathcal{D}$ favor $\mathcal{H}$ if $\mathcal{L}(\mathcal{H}) > 1$, and favor SM if $\mathcal{L}(\mathcal{H}) < 1$. For convenience of interpretation, QUAERO reports $\log_{10}\mathcal{L}(\mathcal{H})$, conveniently thought of as units of "evidence" for or against $\mathcal{H}$.

Systematic errors reduce the evidence the data is able to provide for or against $\mathcal{H}$ relative to SM. In the example of this section, with systematic errors neglected, QUAERO determines

$$\log_{10}\mathcal{L}(\mathcal{H}) = -1.4, \tag{18}$$

providing 1.4 units of evidence against $\mathcal{H}$ relative to the Standard Model. When the systematic errors considered include

- a 5% error on the integrated luminosity,
- a 1% error on the electromagnetic energy scale,
- a 3% error on the hadronic energy scale, and
- a 1% error on electron efficiency,

QUAERO determines

$$\log_{10}\mathcal{L}(\mathcal{H}) = -0.7, \tag{19}$$

providing 0.7 units of evidence against $\mathcal{H}$ relative to the Standard Model. In this way the incorporation of systematic errors leads to greater uncertainty on the measurement of model parameters, as illustrated in Figs. 5 and 6.

QUAERO's sole output is thus a single number — the evidence provided by the data for or against $\mathcal{H}$ relative to SM, after incorporation of systematic errors. The result of an analysis is condensed into 4 bytes.

L. Interpretation of results

For a given hypothesis $\mathcal{H}$, QUAERO's result takes the form of the single number $\mathcal{L}(\mathcal{H})$. In words, $\mathcal{L}(\mathcal{H})$ quantifies the extent to which the data support $\mathcal{H}$ in favor of the Standard Model. If prior prejudice places the betting odds for $\mathcal{H}$ over the Standard Model at $p(\mathcal{H})/p(\mathrm{SM})$, then QUAERO's use of the data $\mathcal{D}$ modifies those odds to

$$\frac{p(\mathcal{H}|\mathcal{D})}{p(\mathrm{SM}|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H})}{p(\mathcal{D}|\mathrm{SM})}\frac{p(\mathcal{H})}{p(\mathrm{SM})}. \tag{20}$$

The new betting odds are obtained from the old simply by multiplication by $\mathcal{L}(\mathcal{H})$.

The likelihood ratio $\mathcal{L}(\mathcal{H})$ can be converted into a more familiar, if perhaps less natural, form. Results in high energy physics are often presented either in terms of a measurement of one or more parameters of a model (central value with one standard deviation errors), or in terms of an exclusion limit for one or more parameters of a model (typically at the 95% confidence level).

In either case the hypothesis $\mathcal{H}$ is taken to depend upon one or more parameters $\vec{\alpha}$. If prior prejudice places the betting odds for $\mathcal{H}(\vec{\alpha})$ over the Standard Model at $p(\mathcal{H}(\vec{\alpha}))/p(\mathrm{SM})$, then QUAERO's result modifies those odds to

$$\frac{p(\mathcal{H}(\vec{\alpha})|\mathcal{D})}{p(\mathrm{SM}|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H}(\vec{\alpha}))}{p(\mathcal{D}|\mathrm{SM})}\frac{p(\mathcal{H}(\vec{\alpha}))}{p(\mathrm{SM})}. \tag{21}$$

The new betting odds are obtained from the old simply by multiplication by $\mathcal{L}(\mathcal{H}(\vec{\alpha}))$. In the example of this section, with $\vec{\alpha} \to M_Z$, if you had been willing to bet \$10 at even odds that $M_Z = 88.0$ GeV (rather than 91.2 GeV) before seeing the data, then after observing these data you should be willing to put up \$10 only if the payoff is \$50 or more.

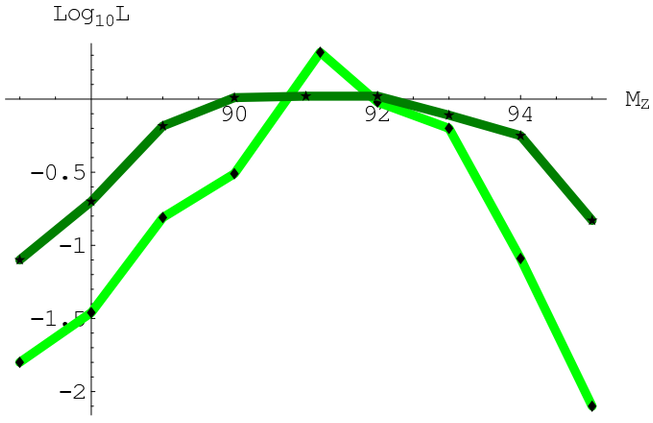

FIG. 5: QUAERO's result $\log_{10}\mathcal{L}$ as a function of the assumed Z boson mass M_Z with systematic errors neglected (light green) and incorporated (dark green). Systematic errors act to reduce the amount evidence provided by the data, flattening the curve. The numerical error on each point is ≈ 0.3.

Using QUAERO to test a number of different hypotheses, each differing from the others only in assuming different values of M_Z, leads to the result shown in Fig. 5. The scale of the numerical error on each point is set by the square root of the weight of the Monte Carlo points; in this example 10 pb^{-1} of Monte Carlo events were generated for 1 pb^{-1} of data, so the numerical error is roughly $0.3 \approx \sqrt{0.1}$. QUAERO's run time increases linearly with the number of Monte Carlo events. This numerical uncertainty is therefore inversely proportional to the square root of the time taken by the algorithm. The example in this section took two CPU minutes on one 1 GHz Linux box.

The posterior distribution $p(M_Z|\mathcal{D})$ obtained from the likelihood in Fig. 5 and a flat prior is shown in Fig. 6. The posterior peaks at the Standard Model value of M_Z, as expected.

The difference between making a measurement, making a discovery, and setting exclusion limits is then loosely as follows. A measurement is being made if $\mathcal{L}(\mathcal{H}(\vec{\alpha}))$ shows a demonstrable peak in $\vec{\alpha}$; a discovery is being made if the hypothesis involves physics beyond the Standard Model and $\mathcal{L}(\mathcal{H}(\vec{\alpha})) \gg 1$; exclusion limits

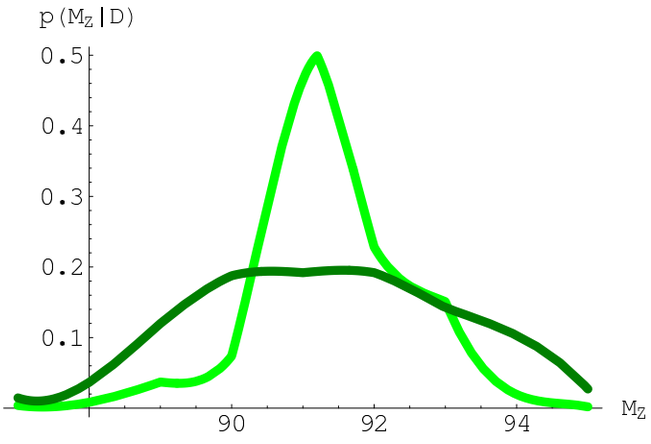

FIG. 6: The posterior distribution $p(M_Z|\mathcal{D})$, obtained from the likelihood in Fig. 5 and a flat prior, with systematic errors neglected (light green) and incorporated (dark green). The broadened posterior distribution reflects the 1% uncertainty on the electromagnetic energy scale and the 5% uncertainty on the integrated luminosity. The posterior peaks at the Standard Model value of M_Z, as expected.

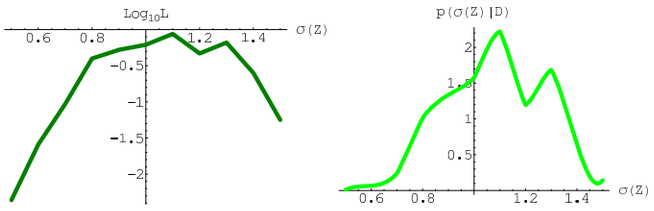

FIG. 7: The posterior distribution $p(\sigma(p\bar{p} \to Z)|\mathcal{D})$ (right), obtained from a flat prior and the likelihood (left) returned by QUAERO. One unit on the horizontal axis corresponds to the Standard Model cross section.

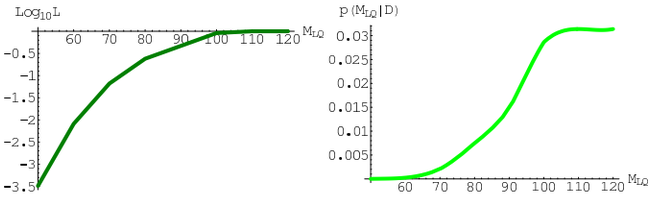

FIG. 8: The posterior distribution $p(M_{LQ}|\mathcal{D})$ (right), obtained from a flat prior and the likelihood (left) returned by QUAERO. Units on the horizontal axis are GeV.

are set in all other cases.

In the case of a measurement, the distribution $p(\mathcal{H}(\vec{\alpha})|\mathcal{D})$ is typically fit to a multivariate gaussian in $\vec{\alpha}$ — the mean of the gaussian then corresponds to the measured central values and the covariance matrix to the errors on those values.

In the case of a discovery, the peak value of $\mathcal{L}(\mathcal{H}(\vec{\alpha}))$ can be quoted directly as a quantitative measure of the traditional "significance" of the result.

In the case of exclusion limits, we typically introduce a cross section σ as a free parameter, ignoring for a moment that the predicted cross section $\sigma(\vec{\alpha})$ is generally a definite function of the parameters $\vec{\alpha}$. The hypothesis $\mathcal{H}(\vec{\alpha})$ is then said to be excluded at the 95% confidence level if

$$\int_0^{\sigma(\vec{\alpha})} p(\mathcal{H}(\vec{\alpha},\sigma)|\mathcal{D})\, d\sigma > 95\%, \qquad (22)$$

assuming some prior $p(\sigma)$ for the cross section.

In all cases, the desired form of the result is easily obtained from the number that QUAERO provides.

III. EXAMPLES

The previous section described the QUAERO algorithm in some detail, using a single example — measuring the Z boson mass — to illustrate each step. This section provides three additional examples.

Figure 7 shows the result of using QUAERO to measure the Z production cross section $\sigma(p\bar{p} \to Z)$ at Tevatron Run II. This is easily accomplished using the pseudo-PYTHIA commands `kfactor=1.1` or `xsec=600` to adjust the level of the signal, and submitting several QUAERO requests varying the values of `kfactor` or `xsec`. One unit of `kfactor` corresponds to the Standard Model cross section (obtained from PYTHIA); one unit of `xsec` corresponds to one picobarn. The expected result — a peak at `kfactor` ≈ 1 is obtained, with reasonable errors. This and subsequent examples incorporate the systematic errors referred to in Sec. II K, and use Monte Carlo corresponding to one hundred times the 1 pb^{-1} of data.

Figure 8 shows a search for leptoquark pair production as a function of assumed leptoquark mass. Leptoquarks with small masses, which would be more copiously produced in the Tevatron than their heavier counterparts, are disfavored by the data. Figure 9 shows the region selected in QUAERO's chosen variable space in the final state $e^+e^-\,2j$ at $m_{LQ} = 50$ GeV. QUAERO separates the SM-enhanced region with small unclustered transverse momentum and $m_{e^+e^-} \approx 91$ GeV from the $\mathcal{H}$-enhanced region with $m_{e^+e^-} < 80$ GeV and large unclustered transverse momentum.

Figure 10 shows a search for a heavy Z' as a function of assumed Z' mass. Z's with small masses, which would be more copiously produced in the Tevatron than their heavier counterparts, are disfavored by the data. The posterior probability $p(m_{Z'}|\mathcal{D})$ flattens out beyond $m_{Z'} \approx 250$ GeV, indicating that the 1 pb^{-1} of data $\mathcal{D}$ is insufficiently sensitive to provide evidence for or against Z's at this mass.

QUAERO's treatment of these toy examples is straightforward, and the results easily understood. The toy-like nature of these examples, although conducive to developing an understanding of the algorithm, precludes a direct assessment of QUAERO's performance relative to its competition. Such an assessment requires the perfor-

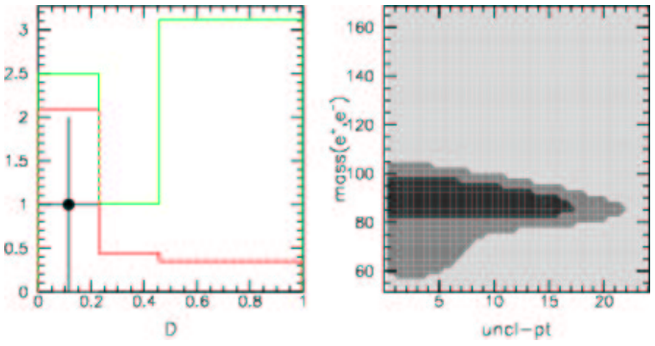

FIG. 9: In the final state e^+e^-j, QUAERO chooses as variables $m_{e^+e^-}$ and the unclustered transverse momentum. Bins in the discriminant D (left) map to regions in the original variable space (right) that separate Standard Model Z production from leptoquark production.

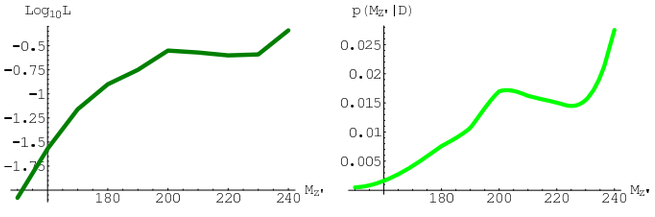

FIG. 10: The posterior distribution $p(M_{Z'}|\mathcal{D})$ (right), obtained from a flat prior and the likelihood (left) returned by QUAERO. Units on the horizontal axis are GeV.

mance of parallel analyses using QUAERO and using a top-notch graduate student, and a direct comparison of results. Work on this front is ongoing.

IV. CONCLUSIONS

QUAERO is a C++ implementation of the analysis algorithm traditionally executed via analysis group meetings, hallway discussions, and hard-working graduate students. The steps of this algorithm, detailed in Sec. II and illustrated through the example of measuring the mass of the Z boson in 1 pb^{-1} of Tevatron data, do not differ greatly from the traditional approach. QUAERO's successful measurement of $\sigma(p\bar{p} \to Z)$ relative to the Standard Model prediction, search for a heavy Z', and search for first generation leptoquarks, described in Sec. III, provide confidence that the algorithm and its implementation in code perform as expected. The procedure described here may prove useful in quickly testing future new physics hypotheses.

APPENDIX A: USE OF DIFFERENT GRADUATE STUDENTS

How do we ensure $\log_{10}(p(\mathcal{D}|\mathcal{H})/p(\mathcal{D}|\mathrm{SM})) = 0$ when $\mathcal{H} = \mathrm{SM}$? The only answer is to use sufficiently large bins. How large is sufficiently large? If we desire $\log_{10}(p(\mathcal{D}|\mathcal{H} = \mathrm{SM})/p(\mathcal{D}|\mathrm{SM})) = 0$ to within $\delta \log_{10} \mathcal{L}$, the maximal weight w in QUAERO's background file must be $w \lesssim (\delta \log_{10} \mathcal{L})^2/10$. The following derives this relation.

Consider a bin with d events observed in the data and h events predicted by hypothesis $\mathcal{H}$. Assume all the Monte Carlo events predicted by $\mathcal{H}$ have equal weight w, so that the bin contains $n = h/w$ individual Monte Carlo events. The question at hand is how much $\log_{10} \mathcal{L}$ varies under changes in the random number seed used to generate these Monte Carlo events. The magnitude (up to factors of order 2) of this variation is roughly given by

$$\delta \log_{10} \mathcal{L} \approx \log_{10} \frac{p(d = h + \sqrt{h}|h + w\sqrt{n})}{p(d = h + \sqrt{h}|h - w\sqrt{n})}. \qquad \text{(A1)}$$

In words, we are considering the case in which the data d has fluctuated one standard deviation $\sqrt{h}$ above the prediction h, and are comparing QUAERO's result in the case that the Monte Carlo estimate has fluctuated one standard deviation low to the case that the Monte Carlo estimate has fluctuated one standard deviation high. The statistical error of the data is $\sqrt{h}$, and the statistical error σ on the Monte Carlo prediction is $\sigma = w\sqrt{n} = h/\sqrt{n}$.

Assuming Gaussian distributions and $w \ll 1$,

$$\begin{aligned} \delta \log_{10} \mathcal{L} &\approx (\log_{10} e) \left(-\frac{(h + \sqrt{h} - (h + w\sqrt{n}))^2}{2(\sqrt{h})^2} + \frac{(h + \sqrt{h} - (h - w\sqrt{n}))^2}{2(\sqrt{h})^2} \right) && \text{(A2)} \\ &= 2(\log_{10} e)\sqrt{w} && \text{(A3)} \\ &\approx \sqrt{w}. && \text{(A4)} \end{aligned}$$

This result is interesting. Several points are worth noting:

- The variation depends upon the total number of predicted events h and the total number of generated Monte Carlo events n only through the weight w of each event.
- A dependence linear in w is obtained if the number of observed data events d is fixed at h, with the

scale of variation given by

$$\delta \log_{10} \mathcal{L} \approx \log_{10} \frac{p(d = h|h)}{p(d = h|h - w\sqrt{n})}. \qquad \text{(A5)}$$

- If k bins are used by QUAERO, the variation under changes of random number seed grows as $\sqrt{k}$.
- The weight w that enters in this derivation is the weight of the events that are used to populate QUAERO's chosen bins, after kernels are estimated and bin edges chosen, which is three times the largest weight appearing in the file provided to QUAERO.

The magnitude of this effect is in agreement with that observed in actual QUAERO requests.

How should this variation in QUAERO's results be understood? It is of the same nature as the variation obtained under choosing different graduate students to perform the same analysis. This is a variation not typically considered or reported. If one hundred graduate students all take an identical hypothesis $\mathcal{H}$ and compute $\delta \log_{10} \mathcal{L}$, there will be a resulting spread of answers.[11] Jane may elect not to use the final state $e\mu\,3j$, Ken may decide to use ${p_T}^e$ and Sandra ${p_T}^\mu$, Burkhard may decide to cut at $\sum p_T > 100$ and Mark at $\sum p_T > 120$. Similarly, depending upon where the Monte Carlo events happen to fall, QUAERO may choose to consider the final state $e\mu\,3j$, to use ${p_T}^e$ rather than ${p_T}^\mu$, and to cut at $\sum p_T > 110$.

QUAERO needs to return numbers accurate to within $\delta \log_{10} \mathcal{L} \approx 0.1$ in order for measurements corresponding to one standard deviation ($\log_{10} \mathcal{L} \approx -0.2$) to be made. Bounding the number of bins considered by QUAERO to three at minimum, and recalling that only the last third of the events is used to populate the chosen bins, the largest weight w presented to QUAERO should be $w < (0.1)^2/3/3 \approx 10^{-3}$.

[1] B. Knuteson (2003), http://mit.fnal.gov/~knuteson/Quaero/quaero/doc/notes/interface.pdf.

[2] B. Knuteson (2004), TURBOSIM: A Self-Tuning Fast Detector Simulation; http://mit.fnal.gov/~knuteson/papers/turboSim.ps.

[3] T. Sjostrand, L. Lonnblad, and S. Mrenna (2001), hep-ph/0108264.

[4] A. Djouadi, J.-L. Kneur, and G. Moultaka (2002), hep-ph/0211331.

[5] F. Maltoni and T. Stelzer (2002), hep-ph/0208156, URL http://madgraph.physics.uiuc.edu/.

[6] J. Conway (1998), http://www.physics.ucdavis.edu/~conway/research/software/pgs/pgs.html.

[7] B. Knuteson (2003), http://mit.fnal.gov/~knuteson/Quaero/quaero/doc/notes/fewKDE.ps.

[8] B. Knuteson (2003), http://mit.fnal.gov/~knuteson/Quaero/quaero/doc/notes/optimalBinning.ps.

[9] $\lfloor \cdot \rfloor$ is the "floor" operator, denoting the largest integer not exceeding its argument.

[10] In this and later figures, the green curve will always show the prediction of $\mathcal{H}$, and the red curve will show the prediction from SM. When viewed in grayscale, the green curve is perceptibly lighter than the red.

[11] Recall here that the $\delta \log_{10} \mathcal{L}$ that each student determines is a *number*, and does not come with an error — systematic and statistical errors have been integrated over, and are all wrapped up in this single value.