| Institution                                                                 | Location                                      |
|----------------------------------------------------------------------------|-----------------------------------------------|
| Institute of High Energy Physics, Beijing, People’s Republic of China      | Beijing, China                                |
| Universidad de los Andes, Bogotá, Colombia                                 | Bogotá, Colombia                              |
| Charles University, Prague, Czech Republic                                 | Prague, Czech Republic                        |
| Institute of Physics, Academy of Sciences, Prague, Czech Republic          | Prague, Czech Republic                        |
| Universidad San Francisco de Quito, Quito, Ecuador                          | Quito, Ecuador                                |
| Institut des Sciences Nucléaires, IN2P3-CNRS, Université de Grenoble 1, Grenoble, France | Grenoble, France                              |
| CPPM, IN2P3-CNRS, Université de la Méditerranée, Marseille, France         | Marseille, France                             |
| LPNHE, Universités Paris VI and VII, IN2P3-CNRS, Paris, France             | Paris, France                                 |
| DAPNIA/Service de Physique des Particules, CEA, Saclay, France             | Saclay, France                                |
| Panjab University, Chandigarh, India                                        | Chandigarh, India                             |
| Delhi University, Delhi, India                                             | Delhi, India                                  |
| Tata Institute of Fundamental Research, Mumbai, India                       | Mumbai, India                                 |
| Seoul National University, Seoul, Korea                                     | Seoul, Korea                                  |
| CINVESTAV, Mexico City, Mexico                                             | Mexico City                                   |
| University of Nijmegen/NIKHEF, Nijmegen, The Netherlands                   | Nijmegen, The Netherlands                     |
| Institute of Nuclear Physics, Kraków, Poland                               | Kraków, Poland                                |
| Institute for Theoretical and Experimental Physics, Moscow, Russia         | Moscow, Russia                                |
| Moscow State University, Moscow, Russia                                     | Moscow, Russia                                |
| Institute for High Energy Physics, Protvino, Russia                        | Protvino, Russia                              |
| Lancaster University, Lancaster, United Kingdom                            | Lancaster, United Kingdom                     |
| University of Arizona, Tucson, Arizona                                     | Tucson, Arizona                               |
| Lawrence Berkeley National Laboratory and University of California, Berkeley, California | Berkeley, California                         |
| University of California, Davis, California                                | Davis, California                             |
| California State University, Fresno, California                             | Fresno, California                            |
| University of California, Irvine, California                               | Irvine, California                            |
| University of California, Riverside, California                             | Riverside, California                         |
| Florida State University, Tallahassee, Florida                              | Tallahassee, Florida                          |
| University of Hawaii, Honolulu, Hawaii                                     | Honolulu, Hawaii                              |
| University of Iowa at Chicago, Chicago, Illinois                           | Chicago, Illinois                             |
| Northern Illinois University, DeKalb, Illinois                             | DeKalb, Illinois                              |
| Northwestern University, Evanston, Illinois                                | Evanston, Illinois                            |
| Indiana University, Bloomington, Indiana                                   | Bloomington, Indiana                          |
| University of Notre Dame, Notre Dame, Indiana                              | Notre Dame, Indiana                           |
| Iowa State University, Ames, Iowa                                         | Ames, Iowa                                    |
| University of Kansas, Lawrence, Kansas                                     | Lawrence, Kansas                              |
| Kansas State University, Manhattan, Kansas                                  | Manhattan, Kansas                             |
| Louisiana Tech University, Ruston, Louisiana                               | Ruston, Louisiana                             |
| University of Maryland, College Park, Maryland                             | College Park, Maryland                        |
| Boston University, Boston, Massachusetts                                   | Boston, Massachusetts                         |
| Northeastern University, Boston, Massachusetts                             | Boston, Massachusetts                         |
| University of Michigan, Ann Arbor, Michigan                                | Ann Arbor, Michigan                           |
| Michigan State University, East Lansing, Michigan                           | East Lansing, Michigan                        |
| University of Nebraska, Lincoln, Nebraska                                  | Lincoln, Nebraska                             |
| Columbia University, New York, New York                                    | New York, New York                            |
| New York University, New York, New York                                   | New York, New York                            |
| University of Rochester, Rochester, New York                               | Rochester, New York                           |
| State University of New York, Stony Brook, New York                        | Stony Brook, New York                         |
| Brookhaven National Laboratory, Upton, New York                             | Upton, New York                               |
| Langston University, Langston, Oklahoma                                    | Langston, Oklahoma                            |
| University of Oklahoma, Norman, Oklahoma                                   | Norman, Oklahoma                              |
| Brown University, Providence, Rhode Island                                 | Providence, Rhode Island                      |
| University of Texas, Arlington, Texas                                      | Arlington, Texas                              |
| Texas A&M University, College Station, Texas                               | College Station, Texas                        |
| Rice University, Houston, Texas                                            | Houston, Texas                                |
| University of Washington, Seattle, Washington                              | Seattle, Washington                           |
Abstract

We present a quasi-model-independent search for the physics responsible for electroweak symmetry breaking. We define final states to be studied, and construct a rule that identifies a set of relevant variables for any particular final state. A new algorithm ("Sleuth") searches for regions of excess in those variables and quantifies the significance of any detected excess. After demonstrating the sensitivity of the method, we apply it to the semi-inclusive channel $e\mu X$ collected in 108 pb$^{-1}$ of $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV at the DØ experiment during 1992–1996 at the Fermilab Tevatron. We find no evidence of new high $p_T$ physics in this sample.
I. INTRODUCTION

It is generally recognized that the standard model, an extremely successful description of the fundamental particles and their interactions, must be incomplete. Although there is likely to be new physics beyond the current picture, the possibilities are sufficiently broad that the first hint could appear in any of many different guises. This suggests the importance of performing searches that are as model-independent as possible.

The word “model” can connote varying degrees of generality. It can mean a particular model together with definite choices of parameters [e.g., mSUGRA] with specified $m_{1/2}$, $m_0$, $A_0$, $\tan \beta$, and $\text{sign}(\mu)$; it can mean a particular model with unspecified parameters (e.g., mSUGRA); it can mean a more general model (e.g., SUGRA); it can mean an even more general model (e.g., gravity-mediated supersymmetry); it can mean a class of general models (e.g., supersymmetry); or it can be a set of classes of general models (e.g., theories of electroweak symmetry breaking). As one ascends this hierarchy of generality, predictions of the “model” become less precise. While there have been many searches for phenomena predicted by models in the narrow sense, there have been relatively few searches for predictions of the more general kind.

In this article we describe an explicit prescription for searching for the physics responsible for stabilizing electroweak symmetry breaking, in a manner that relies only upon what we are sure we know about electroweak symmetry breaking: that its natural scale is on the order of the Higgs mass $\mu$. When we wish to emphasize the generality of the approach, we say that it is quasi-model-independent, where the “quasi” refers to the fact that the correct model of electroweak symmetry breaking should become manifest at the scale of several hundred GeV.

New sources of physics will in general lead to an excess over the expected background in some final state. A general signature for new physics is therefore a region of variable space in which the probability for the background to fluctuate up to or above the number of observed events is small. Because the mass scale of electroweak symmetry breaking is larger than the mass scale of most standard model backgrounds, we expect this excess to populate regions of high transverse momentum ($p_T$). The method we will describe involves a systematic search for such excesses (although with a small modification it is equally applicable to searches for deficits). Although motivated by the problem of electroweak symmetry breaking, this method is generally sensitive to any new high $p_T$ physics.

An important benefit of a precise $a$ priori algorithm of the type we construct is that it allows an $a$ posteriori evaluation of the significance of a small excess, in addition to providing a recipe for searching for such an effect. The potential benefit of this feature can be seen by considering the two curious events seen by the CDF collaboration in their semi-inclusive $\mu$ sample $\mu$ and...
one event in the data sample we analyze in this article, which have prompted efforts to determine the probability that the standard model alone could produce such a result. This is quite difficult to do \textit{a posteriori}, as one is forced to somewhat arbitrarily decide what is meant by “such a result.” The method we describe provides an unbiased and quantitative answer to such questions.

“Sleuth,” a quasi-model-independent prescription for searching for high $p_T$ physics beyond the standard model, has two components:

- the definitions of physical objects and final states, and the variables relevant for each final state; and
- an algorithm that systematically hunts for an excess in the space of those variables, and quantifies the likelihood of any excess found.

We describe the prescription in Secs. I and II. In Sec. I we define the physical objects and final states, and we construct a rule for choosing variables relevant for any final state. In Sec. II we describe an algorithm that searches for a region of excess in a multidimensional space, and determines how unlikely it is that this excess arose simply from a statistical fluctuation, taking account of the fact that the search encompasses many regions of this space. This algorithm is especially useful when applied to a large number of final states. For a first application of Sleuth, we choose the semi-inclusive $e\mu X$ data set ($e\mu X$) because it contains “known” signals (pair production of $W$ bosons and top quarks) that can be used to quantify the sensitivity of the algorithm to new physics, and because this final state is prominent in several models of physics beyond the standard model. In Sec. IV we describe the data set and the expected backgrounds from the standard model and instrumental effects. In Sec. V we demonstrate the sensitivity of the method by ignoring the existence of top quark and $W$ boson pair production, and showing that the method can find these signals in the data. In Sec. VI we apply the Sleuth algorithm to the $e\mu X$ data set assuming the known backgrounds, including $WW$ and $t\bar{t}$, and present the results of a search for new physics beyond the standard model.

II. SEARCH STRATEGY

Most recent searches for new physics have followed a well-defined set of steps: first selecting a model to be tested against the standard model, then finding a measurable prediction of this model that differs as much as possible from the prediction of the standard model, and finally comparing the predictions to data. This is clearly the procedure to follow for a small number of compelling candidate theories. Unfortunately, the resources required to implement this procedure grow almost linearly with the number of theories. Although broadly speaking there are currently only three models with internally consistent methods of electroweak symmetry breaking — supersymmetry, strong dynamics, and theories incorporating large extra dimensions — the number of specific models (and corresponding experimental signatures) is in the hundreds. Of these many specific models, at most one is a correct description of nature.

Another issue is that the results of searches for new physics can be unintentionally biased because the number of events under consideration is small, and the details of the analysis are often not specified before the data are examined. An \textit{a priori} technique would permit a detailed study without fear of biasing the result.

We first specify the prescription in a form that should be applicable to any collider experiment sensitive to physics at the electroweak scale. We then provide aspects of the prescription that are specific to DØ. Other experiments wishing to use this prescription would specify similar details appropriate to their detectors.

A. General prescription

We begin by defining final states, and follow by motivating the variables we choose to consider for each of those final states. We assume that standard particle identification requirements, often detector-specific, have been agreed upon. The understanding of all backgrounds, through Monte Carlo programs and data, is crucial to this analysis, and requires great attention to detail. Standard methods for understanding backgrounds — comparing different Monte Carlos, normalizing background predictions to observation, obtaining instrumental backgrounds from related samples, demonstrating agreement in limited regions of variable space, and calibrating against known physical quantities, among many others — are needed and used in this analysis as in any other. Uncertainties in backgrounds, which can limit the sensitivity of the search, are naturally folded into this approach.

1. Final states

In this subsection we partition the data into final states. The specification is based on the notions of exclusive channels and standard particle identification.

\textit{a. Exclusiveness.} Although analyses are frequently performed on inclusive samples, considering only exclusive final states has several advantages in the context of this approach:

- the presence of an extra object (electron, photon, muon, . . . ) in an event often qualitatively affects the probable interpretation of the event;
- the presence of an extra object often changes the variables that are chosen to characterize the final state; and

...
• using inclusive final states can lead to ambiguities when different channels are combined.

We choose to partition the data into exclusive categories.

b. Particle identification. We now specify the labeling of these exclusive final states. The general principle is that we label the event as completely as possible, as long as we have a high degree of confidence in the label. This leads naturally to an explicit prescription for labeling final states.

Most multipurpose experiments are able to identify electrons, muons, photons, and jets, and so we begin by considering a final state to be described by the number of isolated electrons, muons, photons, and jets observed in the event, and whether there is a significant imbalance in transverse momentum \((p_{T})\). We treat \(p_{T}\) as an object in its own right, which must pass certain quality criteria. If \(b\)-tagging, \(c\)-tagging, or \(\tau\)-tagging is possible, then we can differentiate among jets arising from \(b\) quarks, \(c\) quarks, light quarks, and hadronic tau decays. If a magnetic field can be used to obtain the electric charge of a lepton, we can be used to obtain the electric charge of a lepton, we can split the charged leptons \(\ell\) into \(\ell^+\) and \(\ell^-\) but consider final states that are related through global charge conjugation to be equivalent in \(p\bar{p}\) or \(e^+e^-\) (but not \(pp\)) collisions. Thus \(e^+e^-\gamma\) is a different final state than \(e^+\gamma\) and \(e^-\gamma\) together make up a single final state. The definitions of these objects are logically specified for general use in all analyses, and we use these standard identification criteria to define our objects.

We can further specify a final state by identifying any \(W\) or \(Z\) bosons in the event. This has the effect (for example) of splitting the \(eejj\), \(\mu\mu jj\), and \(\tau\tau jj\) final states into the \(Zjj\), \(eejj\), \(\mu\mu jj\), and \(\tau\tau jj\) channels, and splitting the \(eE_Tjj\), \(\mu E_Tjj\), and \(\tau E_Tjj\) final states into \(Wjj\), \(eE_Tjj\), \(\mu E_Tjj\), and \(\tau E_Tjj\) channels. We combine a \(\ell^+\ell^-\) pair into a \(Z\) if their invariant mass \(M_{\ell^+\ell^-}\) falls within a \(Z\) boson mass window (82 \(\leq M_{\ell^+\ell^-} \leq 100\) GeV for DO data) and the event contains neither significant \(E_T\) nor a third charged lepton. If the event contains exactly one photon in addition to a \(\ell^+\ell^-\) pair and contains neither significant \(E_T\) nor a third charged lepton, and if \(M_{\ell^+\ell^-}\) does not fall within the \(Z\) boson mass window, but \(M_{\ell^+\ell^-}\) does, then the \(\ell^+\ell^-\) triplet becomes a \(Z\) boson. If the experiment is not capable of distinguishing between \(\ell^+\) and \(\ell^-\) and the event contains exactly two \(\ell\)'s, they are assumed to have opposite charge. A lepton and \(E_T\) become a \(W\) boson if the transverse mass \(M_T^{E_T}\) is within a \(W\) boson mass window (30 \(\leq M_T^{E_T} \leq 110\) GeV for DO data) and the event contains no second charged lepton. Because the \(W\) boson mass window is so much wider than the \(Z\) boson mass window, we make no attempt to identify radiative \(W\) boson decays.

We do not identify top quarks, gluons, nor \(W\) or \(Z\) bosons from hadronic decays because we would have little confidence in such a label. Since the predicted cross sections for new physics are comparable to those for the production of detectable \(ZZ\), \(WZ\), and \(WW\) final states, we also elect not to identify these final states.

c. Choice of final states to study. Because it is not realistic to specify backgrounds for all possible exclusive final states, choosing prospective final states is an important issue. Theories of physics beyond the standard model make such wide-ranging predictions that neglect of any particular final state purely on theoretical grounds would seem unwise. Focusing on final states in which the data themselves suggest something interesting can be done without fear of bias if all final states and variables for those final states are defined prior to examining the data. Choosing variables is the subject of the next section.

2. Variables

We construct a mapping from each final state to a list of key variables for that final state using a simple, well-motivated, and short set of rules. The rules, which are summarized in Table 1, are obtained through the following reasoning:

- There is strong reason to believe that the physics responsible for electroweak symmetry breaking occurs at the scale of the mass of the Higgs boson, or on the order of a few hundred GeV. Any new massive particles associated with this physics can therefore be expected to decay into objects with large transverse momenta in the final state.

- Many models of electroweak symmetry breaking predict final states with large missing transverse energy. This arises in a large class of \(R\)-parity conserving supersymmetric theories containing a neutral, stable, lightest supersymmetric particle: in theories with “large” extra dimensions containing a Kaluza-Klein tower of gravitons that escape into the multidimensional “bulk space” \(\mathbb{R}^4\), and more generally from neutrinos produced in electroweak boson decay. If the final state contains significant \(E_T\), then \(E_T\) is included in the list of promising variables. We do not use \(E_T\) that is reconstructed as a \(W\) boson decay product, following the prescription for \(W\) and \(Z\) boson identification outlined above.

- If the final state contains one or more leptons we use the summed scalar transverse momenta \(\sum p_T\), where the sum is over all leptons whose identity can be determined and whose momenta can be accurately measured. Leptons that are reconstructed as \(W\) or \(Z\) boson decay products are not included in this sum, again following the prescription for \(W\) and \(Z\) boson identification outlined above. We combine the momenta of \(e, \mu\), and \(\tau\) leptons because these objects are expected to have comparable transverse momenta on the basis of lepton uni-
versatility in the standard model and the negligible values of lepton masses.

- Similarly, photons and $W$ and $Z$ bosons are most likely to signal the presence of new phenomena when they are produced at high transverse momentum. Since the expected transverse momenta of the electroweak gauge bosons are comparable, we use the variable $\sum p_T^{\gamma/W/Z}$, where the scalar sum is over all electroweak gauge bosons in the event, for final states with one or more of them identified.

- For events with one jet in the final state, the transverse energy of that jet is an important variable. For events with two or more jets in the final state, previous analyses have made use of the sum of the transverse energies of all but the leading jet. The reason for excluding the energy of the leading jet from this sum is that while a hard jet is often obtained from QCD radiation, hard second and third radiative jets are relatively much less likely. We therefore choose the variable $\sum p_T^j$ to describe the jets in the final state, where $\sum p_T^j$ denotes $p_T^j$ if the final state contains only one jet, and $\sum_{i=2}^n p_T^j$ if the final state contains two or more jets. Since QCD dijets are a large background in all-jets final states, $\sum p_T^j$ refers instead to $\sum_{i=3}^n p_T^j$ for final states containing $n$ jets and nothing else, where $n \geq 3$.

When there are exactly two objects in an event (e.g., one $Z$ boson and one jet), their $p_T$ values are expected to be nearly equal, and we therefore use the average $p_T$ of the two objects. When there is only one object in an event (e.g., a single $W$ boson), we use no variables, and simply perform a counting experiment.

Other variables that can help pick out specific signatures can also be defined. Although variables such as invariant mass, angular separation between particular final state objects, and variables that characterize event topologies may be useful in testing a particular model, these variables tend to be less powerful in a general search. Appendix contains a more detailed discussion of this point. In the interest of keeping the list of variables as general, well-motivated, powerful, and short as possible, we elect to stop with those given in Table I. We expect evidence for new physics to appear in the high tails of the $E_T$, $\sum p_T^\ell$, $\sum p_T^\gamma/W/Z$, and $\sum p_T^j$ distributions.

### B. Search strategy: DØ Run I

The general search strategy just outlined is applicable to any collider experiment searching for the physics responsible for electroweak symmetry breaking. Any particular experiment that wishes to use this strategy needs to specify object and variable definitions that reflect the capabilities of the detector. This section serves this function for the DØ detector in its 1992–1996 run (Run I) at the Fermilab Tevatron. Details in this subsection supersede those in the more general section above.

#### 1. Object definitions

The particle identification algorithms used here for electrons, muons, jets, and photons are similar to those used in many published DØ analyses. We summarize them here.

- **Electrons.** DØ had no central magnetic field in Run I; therefore, there is no way to distinguish between electrons and positrons. Electron candidates with transverse energy greater than 15 GeV, within the fiducial region of $\eta > 1.1$ or $1.5 < \eta < 2.5$ (where $\eta = -\ln \tan(\theta/2)$, with $\theta$ the polar angle with respect to the colliding proton’s direction), and satisfying standard electron identification and isolation requirements as defined in Ref. are accepted.

- **Muons.** We do not distinguish between positively and negatively charged muons in this analysis. We accept muons with transverse momentum greater than 15 GeV and $|\eta| < 1.7$ that satisfy standard muon identification and isolation requirements.

- **$E_T$.** The missing transverse energy, $E_T$, is the energy required to balance the measured energy in the event. In the calorimeter, we calculate

$$E_T^{\text{cal}} = \sum E_i \sin \theta_i (\cos \phi_i \hat{x} + \sin \phi_i \hat{y}),$$

where $i$ runs over all calorimeter cells, $E_i$ is the energy deposited in the $i^{th}$ cell, and $\phi_i$ and $\theta_i$ are the polar angle of the center of the $i^{th}$ cell, measured with respect to the event vertex.

An event is defined to contain a $E_T$ “object” only if we are confident that there is significant missing transverse
energy. Events that do not contain muons are said to contain $E_T$ if $E_T^{\text{cal}}>15$ GeV. Using track deflection in magnetized steel toroids, the muon momentum resolution in Run I is

$$\delta(1/p) = 0.18(p-2)/p^2 \oplus 0.003,$$

where $p$ is in units of GeV, and the $\oplus$ means addition in quadrature. This is significantly coarser than the electromagnetic and jet energy resolutions, parameterized by

$$\delta E/E = 15\%/\sqrt{E} \oplus 0.3\%$$

and

$$\delta E/E = 80\%/\sqrt{E},$$

respectively. Events that contain exactly one muon are deemed to contain $E_T$ on the basis of muon number conservation rather than on the basis of the muon momentum measurement. We do not identify a $E_T$ object in events that contain two or more muons.

d. Jets. Jets are reconstructed in the calorimeter using a fixed-size cone algorithm, with a cone size of $\Delta R = \sqrt{(\Delta \phi)^2 + (\Delta \eta)^2} = 0.5$. We require jets to have $E_T > 15$ GeV and $|\eta|<2.5$. We make no attempt to distinguish among light quarks, gluons, charm quarks, bottom quarks, and hadronic tau decays.

e. Photons. Isolated photons that pass standard identification requirements [14], have transverse energy greater than 15 GeV, and are in the fiducial region $|\eta|<1.1$ or $1.5<|\eta|<2.5$ are labeled photon objects.
f. W bosons. Following the general prescription described above, an electron (as defined above) and $E_T$ become a $W$ boson if their transverse mass is within the $W$ boson mass window ($30 \leq M_T^{\gamma}\leq 110$ GeV), and the event contains no second charged lepton. Because the muon momentum measurement is coarse, we do not use a transverse mass window for muons. From Sec. [13] any event containing a single muon is said to also contain $E_T$; thus any event containing a muon and no second charged lepton is said to contain a $W$ boson.

g. Z bosons. We use the rules in the previous section for combining an $ee$ pair or $ee\gamma$ triplet into a $Z$ boson. We do not attempt to reconstruct a $Z$ boson in events containing three or more charged leptons. For events containing two muons and no third charged lepton, we fit the event to the hypothesis that the two muons are decay products of a $Z$ boson and that there is no $E_T$ in the event. If the fit is acceptable, the two muons are considered to be a $Z$ boson.

2. Variables

The variables provided in the general prescription above also need minor revision to be appropriate for the DØ experiment.

- $a. \sum p_T^\ell$. We do not attempt to identify $\tau$ leptons, and the momentum resolution for muons is coarse. For events that contain no leptons other than muons, we define $\sum p_T^\ell = \sum p_T^\nu$. For events that contain one or more electrons, we define $\sum p_T^\ell = \sum p_T^e$. This is identical to the general definition provided above except for events containing both one or more electrons and one or more muons. In this case, we have decided to define $\sum p_T^\ell$ as the sum of the momenta of the electrons only, rather than combining the well-measured electron momenta with the poorly-measured muon momenta.

- $b. E_T$. $E_T$ is defined by $E_T = E_T^{\text{cal}}$, where $E_T^{\text{cal}}$ is the missing transverse energy as summed in the calorimeter. This sum includes the $p_T$ of electrons, but only a negligible fraction of the $p_T$ of muons.

- $c. \sum p_T^{\gamma/W/Z}$. We use the definition of $\sum p_T^{\gamma/W/Z}$ provided in the general prescription: the sum is over all electroweak gauge bosons in the event, for final states with one or more of them. We note that if a $W$ boson is formed from a $\mu$ and $E_T$, then $p_T^W = E_T^{\text{cal}}$.

III. SLEUTH ALGORITHM

Given a data sample, its final state, and a set of variables appropriate to that final state, we now describe the algorithm that determines the most interesting region in those variables and quantifies the degree of interest.

A. Overview

Central to the algorithm is the notion of a “region” ($R$). A region can be regarded simply as a volume in the variable space defined by Table [13], satisfying certain special properties to be discussed in Sec. [11]. The region contains $N$ data points and an expected number of background events $b_R$. We can consequently compute the weighted probability $p_R^N$, defined in Sec. [11], that the background in the region fluctuates up to or beyond the observed number of events. If this probability is small, we flag the region as potentially interesting.

In any reasonably-sized data set, there will always be regions in which the probability for $b_R$ to fluctuate up to or above the observed number of events is small. The relevant issue is how often this can happen in an ensemble of hypothetical similar experiments (hse’s). This question can be answered by performing these hypothetical similar experiments; i.e., by generating random events drawn from the background distribution, finding the least probable region, and repeating this many times. The fraction of hypothetical similar experiments that yields a probability as low as the one observed in the data provides the appropriate measure of the degree of interest.

Although the details of the algorithm are complex, the interface is straightforward. What is needed is a data sample, a set of events for each background process $i$,
and the number of background events \( \hat{b}_i \pm \delta b_i \) from each background process expected in the data sample. The output gives the region of greatest excess and the fraction of hypothetical similar experiments that would yield such an excess.

The algorithm consists of seven steps:

1. Define regions \( R \) about any chosen set of \( N = 1, \ldots, N_{\text{data}} \) data points in the sample of \( N_{\text{data}} \) data points.
2. Estimate the background \( \hat{b}_R \) expected within these \( R \).
3. Calculate the weighted probabilities \( p^R_N \) that \( b_R \) can fluctuate to \( \geq N \).
4. For each \( N \), determine the \( R \) for which \( p^R_N \) is minimum. Define \( p_N = \min_R (p^R_N) \).
5. Determine the fraction \( P_N \) of hypothetical similar experiments in which the \( p_N(\text{hse}) \) is smaller than the observed \( p_N(\text{data}) \).
6. Determine the \( N \) for which \( P_N \) is minimized. Define \( P = \min_N (P_N) \).
7. Determine the fraction \( \mathcal{P} \) of hypothetical similar experiments in which the \( P(\text{hse}) \) is smaller than the observed \( P(\text{data}) \).

Our notation is such that a lowercase \( p \) represents a probability, while an uppercase \( P \) or \( \mathcal{P} \) represents the fraction of hypothetical similar experiments that would yield a less probable outcome. The symbol representing the minimization of \( p^R_N \) over \( R \), \( p_N \) over \( N \), or \( P_N \) over \( N \) is written without the superscript or subscript representing the varied property (i.e., \( p_N, p, \) or \( P, \) respectively). The rest of this section discusses these steps in greater detail.

### B. Steps 1 and 2: Regions

When there are events that do not appear to follow some expected distribution, such as the event at \( x = 61 \) in Fig. 1, we often attempt to estimate the probability that the event is consistent with coming from that distribution. This is generally done by choosing some region around the event (or an accumulation of events), integrating the background within that region, and computing the probability that the expected number of events in that region could have fluctuated up to or beyond the observed number.

Of course, the calculated probability depends on how the region containing the events is chosen. If the region about the event is infinitesimal, then the expected number of background events in the region (and therefore this probability) can be made arbitrarily small. A possible approach in one dimension is to define the region to be the interval bounded below by the point halfway between the interesting event and its nearest neighbor, and bounded above by infinity. For the case shown in Fig. 1, this region would be roughly the interval \((46, \infty)\).

Such a prescription breaks down in two or more dimensions, and it is not entirely satisfactory even in one dimension. In particular, it is not clear how to proceed if the excess occurs somewhere other than at the tail end of a distribution, or how to generalize the interval to a well-defined contour in several dimensions. As we will see, there are significant advantages to having a precise definition of a region about a potentially interesting set of data points. This is provided in Sec. III B 2, after we specify the variable space itself.

#### 1. Variable transformation

Unfortunately, the region that we choose about the point on the tail of Fig. 1 changes if the variable is some function of \( x \), rather than \( x \) itself. If the region about each data point is to be the subspace that is closer to that point than to any other one in the sample, it would therefore be wise to minimize any dependence of the selection on the shape of the background distribution. For a background distributed uniformly between 0 and 1 (or, in \( d \) dimensions, uniform within the unit “box” \([0, 1]^d\)), it is reasonable to define the region associated with an event as the variable subspace closer to that event than to any other event in the sample. If the background is not already uniform within the unit box, we transform
the variables so that it becomes uniform. The details of this transformation are provided in Appendix B.

With the background distribution trivialized, the rest of the analysis can be performed within the unit box without worrying about the background shape. A considerable simplification is therefore achieved through this transformation. The task of determining the expected background within each region, which would have required a Monte Carlo integration of the background distribution over the region, reduces to the problem of determining the volume of each region. The problem is now completely specified by the transformed coordinates of the data points, the total number of expected background events $b$, and its uncertainty $\delta b$.

2. Voronoi diagrams

Having defined the variable space by requiring a uniform background distribution, we can now define more precisely what is meant by a region. Figure 3 shows a 2-dimensional variable space $V$ containing seven data points in a unit square. For any $v \in V$, we say that $v$ belongs to the data point $D_i$ if $\lVert v - D_i \rVert < \lVert v - D_j \rVert$ for all $j \neq i$; that is, $v$ belongs to $D_i$ if $v$ is closer to $D_i$ than to any other data point. In Fig. 3(a), for example, any $v$ lying within the variable subspace defined by the pentagon in the upper right-hand corner belongs to the data point located at $(0, 0.9)$. The set of points in $V$ that do not belong to any data point [those points on the lines in Fig. 3(a)] has zero measure and may be ignored.

We define a region around a set of data points in a variable space $V$ to be the set of all points in $V$ that are closer to one of the data points in that set than to any data points outside that set. A region around a single data point is the union of all points in $V$ that belong to that data point, and is called a 1-region. A region about a set of $N$ data points is the union of all points in $V$ that belong to any one of the data points, and is called an $N$-region; an example of a 2-region is shown as the shaded area in Fig. 3(b). $N_{\text{data}}$ data points thus partition $V$ into $N_{\text{data}}$ 1-regions. Two data points are said to be neighbors if their 1-regions share a border – the points at $(0.75, 0.9)$ and $(0.9, 0.8)$ in Fig. 3, for example, are neighbors. A diagram such as Fig. 3(a), showing a set of data points and their regions, is known as a Voronoi diagram. We use a program called HULL [13] for this computation.

3. Region criteria

The explicit definition of a region that we have just provided reduces the number of contours we can draw in the variable space from infinite to a mere $2^{N_{\text{data}}} - 1$, since any region either contains all of the points belonging to the $i^{th}$ data event or it contains none of them. In fact, because many of these regions have a shape that makes them implausible as “discovery regions” in which new physics might be concentrated, the number of possible regions may be reduced further. For example, the region in Fig. 3 containing only the lower-leftmost and the upper-rightmost data points is unlikely to be a discovery region, whereas the region shown in Fig. 3(b) containing the two upper-rightmost data points is more likely (depending upon the nature of the variables).

We can now impose whatever criteria we wish upon the regions that we allow Sleuth to consider. In general we will want to impose several criteria, and in this case we write the net criterion $c_R = c^{\text{int}}_R c^{\text{out}}_R \ldots$ as a product of the individual criteria, where $c^{\text{int}}_R$ is to be read “the extent to which the region $R$ satisfies the criterion $c^i$.” The quantities $c^{\text{int}}_R$ take on values in the interval $[0, 1]$, where $c^{\text{int}}_R \to 0$ if $R$ badly fails $c^i$, and $c^{\text{int}}_R \to 1$ if $R$ easily satisfies $c^i$.

Consider as an example $c = \text{AntiCornerSphere}$, a simple criterion that we have elected to impose on the regions in the $e\mu X$ sample. Loosely speaking, a region $R$ will satisfy this criterion ($c_R \to 1$) if all of the data points inside the region are farther from the origin than all of the data points outside the region. This situation is shown, for example, in Fig. 3(b). For every event $i$ in the data set, denote by $r_i$ the distance of the point in the unit box to the origin, let $r'$ be $r$ transformed so that the background is uniform in $r'$ over the interval $[0, 1]$, and let $r'^i$ be the values $r_i$ so transformed. Then define

$$c_R = \begin{cases} 0 & 0 \leq \frac{1}{2} + \frac{r'^{\text{int}}_R - r'^{\text{out}}_R}{\xi} \leq 1 \\ \frac{1}{2} + \frac{r'^{\text{int}}_R - r'^{\text{out}}_R}{\xi} & 1 < \frac{1}{2} + \frac{r'^{\text{int}}_R - r'^{\text{out}}_R}{\xi} \end{cases}$$

where $r'^{\text{int}}_R = \min_{i \in R} (r'^i)$, $r'^{\text{out}}_R = \max_{i \notin R} (r'^i)$, and $\xi = 1/(4N_{\text{data}})$ is an average separation distance between data points in the variable $r'$.

Notice that in the limit of vanishing $\xi$, the criterion $c$
becomes a boolean operator, returning “true” when all of the data points inside the region are farther from the origin than all of the data points outside the region, and “false” otherwise. In fact, many possible criteria have a scale $\xi$ and reduce to boolean operators when $\xi$ vanishes. This scale has been introduced to ensure continuity of the final result under small changes in the background estimate. In this spirit, the “extent to which $R$ satisfies the criterion $c$” has an alternative interpretation as the “fraction of the time $R$ satisfies the criterion $c$,” where the average is taken over an ensemble of slightly perturbed background estimates and $\xi$ is taken to vanish, so that “satisfies” makes sense. We will use $c_R$ in the next section to define an initial measure of the degree to which $R$ is interesting.

We have considered several other criteria that could be imposed upon any potential discovery region to ensure that the region is “reasonably shaped” and “in a believable location.” We discuss a few of these criteria in Appendix C.

C. Step 3: Probabilities and uncertainties

Now that we have specified the notion of a region, we can define a quantitative measure of the “degree of interest” of a region.

1. Probabilities

Since we are looking for regions of excess, the appropriate measure of the degree of interest is a slight modification of the probability of background fluctuating up to or above the observed number of events. For an $N$-region $R$ in which $\hat{b}_R$ background events are expected and $\hat{b}_R$ is precisely known, this probability is

$$\sum_{i=N}^{\infty} \frac{e^{-\hat{b}_R}(\hat{b}_R)^i}{i!}. \tag{6}$$

We use this to define the weighted probability

$$p^R_N = \left(\sum_{i=N}^{\infty} \frac{e^{-\hat{b}_R}(\hat{b}_R)^i}{i!}\right) c_R + (1 - c_R), \tag{7}$$

which one can also think of as an “average probability,” where the average is taken over the ensemble of slightly perturbed background estimates referred to above. By construction, this quantity has all of the properties we need: it reduces to the probability in Eq. 6 in the limit that $R$ easily satisfies the region criteria, it saturates at unity in the limit that $R$ badly fails the region criteria, and it exhibits continuous behavior under small perturbations in the background estimate between these two extremes.

2. Systematic uncertainties

The expected number of events from each background process has a systematic uncertainty that must be taken into account. There may also be an uncertainty in the shape of a particular background distribution — for example, the tail of a distribution may have a larger systematic uncertainty than the mode.

The background distribution comprises one or more contributing background processes. For each background process we know the number of expected events and the systematic uncertainty on this number, and we have a set of Monte Carlo points that tell us what that background process looks like in the variables of interest. A typical situation is sketched in Fig. 3.

FIG. 3. An example of a one-dimensional background distribution with three sources. The normalized shapes of the individual background processes are shown as the dashed lines; the solid line is their sum. Typically, the normalizations for the background processes have separate systematic errors. These errors can change the shape of the total background curve in addition to its overall normalization. For example, if the long-dashed curve has a large systematic error, then the solid curve will be known less precisely in the region $(3,5)$ than in the region $(0,3)$ where the other two backgrounds dominate.

The multivariate transformation described in Sec. II B is obtained assuming that the number of events expected from each background process is known precisely. This fixes each event’s position in the unit box, its neighbors, and the volume of the surrounding region. The systematic uncertainty $\delta b_R$ on the number of background events in a given region is computed by combining the systematic uncertainties for each individual background process. Eq. 7 then generalizes to

$$p^R_N = c_R \int_0^{\infty} \sum_{i=N}^{\infty} \frac{e^{-b}(\frac{1}{\sqrt{2\pi(\delta b_R^2)}} \times \exp \left(-\frac{(b - \hat{b}_R)^2}{2(\delta b_R)^2}\right) db}{i!} \times$$

$$+ (1 - c_R), \tag{8}$$
which is seen to reduce to Eq. (3) in the limit $\delta b_R \to 0$.

This formulation provides a way to take account of systematic uncertainties on the shapes of distributions, as well. For example, if there is a larger systematic uncertainty on the tail of a distribution, then the background process can be broken into two components, one describing
the bulk of the distribution and one describing the tail, and a larger systematic uncertainty assigned to the piece that describes the tail. Correlations among the various components may also be assigned.

We vary the number of events generated in the hypothetical similar experiments according to the systematic and statistical uncertainties. The systematic errors are accounted for by pulling a vector of the "true" number of expected background events $\hat{b}$ from the distribution

$$p(\hat{b}) = \frac{1}{\sqrt{2\pi |\Sigma|}} \exp \left( -\frac{1}{2} (b_i - \hat{b}_i) \Sigma^{-1} (b_j - \hat{b}_j) \right), \quad (9)$$

where $\hat{b}_i$ is the $i^{th}$ component of the vector $\hat{b}$. We have introduced a covariance matrix $\Sigma$, which is diagonal with components $\Sigma_{ii} = (\delta b_i)^2$ in the limit that the systematic uncertainties on the different background processes are uncorrelated, and we assume summation on repeated indices in Eq. (9). The statistical uncertainties in turn are allowed for by choosing the number of events $N_i$ from each background process $i$ from the Poisson distribution

$$P(N_i) = \frac{e^{-b_i} b_i^{N_i}}{N_i!}. \quad (10)$$

where $b_i$ is the $i^{th}$ component of the vector $\hat{b}$ just determined.

**D. Step 4: Exploration of regions**

Knowing how to calculate $p_N^{R}$ for a specific $N$-region $R$ allows us to determine which of two $N$-regions is more interesting. Specifically, an $N$-region $R_1$ is more interesting than another $N$-region $R_2$ if $p_N^{R_1} < p_N^{R_2}$. This allows us to compare regions of the same size (the same $N$), although, as we will see, it does not allow us to compare regions of different size.

Step 4 of the algorithm involves finding the most interesting $N$-region for each fixed $N$ between 1 and $N_{data}$. This most interesting $N$-region is the one that minimizes $p_N^{R}$, and these $p_N = \min_R p_N^{R}$ are needed for the next step in the algorithm.

Even for modestly sized problems (say, two dimensions with the order of 100 data points), there are far too many regions to consider an exhaustive search. We therefore use a heuristic to find the most interesting region. We imagine the region under consideration to be an amoeba moving within the unit box. At each step in the search the amoeba either expands or contracts according to certain rules, and along the way we keep track of the most interesting $N$-region so far found, for each $N$. The detailed rules for this heuristic are provided in Appendix E.

**E. Steps 5 and 6: Hypothetical similar experiments, Part I**

At this point in the algorithm the original events have been reduced to $N_{data}$ values, each between 0 and 1: the $p_N$ ($N = 1, \ldots, N_{data}$) corresponding to the most interesting $N$-regions satisfying the imposed criteria. To find the most interesting of these, we need a way of comparing regions of different size (different $N$). An $N_1$-region $R_{hse}$ with $p_{N_1}^{data}$ is more interesting than an $N_2$-region $R_{hse}$ with $p_{N_2}^{data}$ if the fraction of hypothetical similar experiments in which $p_{N_1}^{hse} < p_{N_1}^{data}$ is less than the fraction of hypothetical similar experiments in which $p_{N_2}^{hse} < p_{N_2}^{data}$.

To make this comparison, we generate $N_{hse}$ hypothetical similar experiments. Generating a hypothetical similar experiment involves pulling a random integer from Eq. (9) for each background process $i$, sampling this number of events from the multidimensional background density $b(\hat{x})$, and then transforming these events into the unit box.

For each $hse$ we compute a list of $p_N$, exactly as for the data set. Each of the $N_{hse}$ hypothetical similar experiments consequently yields a list of $p_N$. For each $N$, we now compare the $p_N$ we obtained in the data ($p_{N_{data}}^{data}$) with the $p_N$’s we obtained in the $hse$’s ($p_{N_{hse}}^{hse}$, where $i = 1, \ldots, N_{hse}$). From these values we calculate $P_N$, the fraction of $hse$’s with $p_{N_{hse}}^{hse} < p_{N_{data}}^{data}$:

$$P_N = \frac{1}{N_{hse}} \sum_{i=1}^{N_{hse}} \Theta \left( p_{N_{data}}^{data} - p_{N_{hse}}^{hse} \right), \quad (11)$$

where $\Theta(x) = 0$ for $x < 0$, and $\Theta(x) = 1$ for $x \geq 0$.

The most interesting region in the sample is then the region for which $P_N$ is smallest. We define $P = P_{N_{min}}$, where $P_{N_{min}}$ is the smallest of the $P_N$.

**F. Step 7: Hypothetical similar experiments, Part II**

A question that remains to be answered is what fraction $P$ of hypothetical similar experiments would yield a $P$ less than the $P$ obtained in the data. We calculate $P$ by running a second set of $N_{hse}$ hypothetical similar experiments, generated as described in the previous section. (We have written $hse$ above to refer to the first set of hypothetical similar experiments, used to determine the $P_N$, given a list of $p_N$; we write $hse$’s to refer to this second set of hypothetical similar experiments, used to determine $P$ from $P$.) A second, independent set of $hse$’s...
is required to calculate an unbiased value for $\mathcal{P}$. The quantity $\mathcal{P}$ is then given by

$$\mathcal{P} = \frac{1}{N_{\text{hse}}^2} \sum_{i=1}^{N_{\text{hse}}} \Theta(p_{\text{data}} - p_{\text{hse}}^2).$$  \hspace{1cm} (12)

This is the final measure of the degree of interest of the most interesting region. Note that $\mathcal{P}$ is a number between 0 and 1, that small values of $\mathcal{P}$ indicate a sample containing an interesting region, that large values of $\mathcal{P}$ indicate a sample containing no interesting region, and that $\mathcal{P}$ can be described as the fraction of hypothetical similar experiments that yield a more interesting result than is observed in the data. $\mathcal{P}$ can be translated into units of standard deviations ($\mathcal{P}_{[\sigma]}$) by solving the unit conversion equation

$$\mathcal{P} = \frac{1}{\sqrt{2\pi}} \int_{\mathcal{P}_{[\sigma]}}^{\infty} e^{-t^2/2} dt$$  \hspace{1cm} (13)

for $\mathcal{P}_{[\sigma]}$.

G. Interpretation of results

In a general search for new phenomena, Sleuth will be applied to $N_{fs}$ different final states, resulting in $N_{fs}$ different values for $\mathcal{P}$. The final step in the procedure is the combination of these results. If no $\mathcal{P}$ value is smaller than $\approx 0.01$ then a null result has been obtained, as no significant signal for new physics has been identified in the data.

If one or more of the $\mathcal{P}$ values is particularly low, then we can surmise that the region(s) of excess corresponds either to a poorly modeled background or to possible evidence of new physics. The algorithm has pointed out a region of excess ($\mathcal{R}$) and has quantified its significance ($\mathcal{P}$). The next step is to interpret this result.

Two issues related to this interpretation are combining results from many final states, and confirming a Sleuth discovery.

1. Combining the results of many final states

If one looks at many final states, one expects eventually to see a fairly small $\mathcal{P}$, even if there really is no new physics in the data. We therefore define a quantity $\mathcal{P}$ to be the fraction of hypothetical similar experimental runs that yield a $\mathcal{P}$ that is smaller than the smallest $\mathcal{P}$ observed in the data. Explicitly, given $N_{fs}$ final states, with $b_i$ background events expected in each, and $\mathcal{P}_i$ calculated for each one, $\mathcal{P}$ is given to good approximation by

$$\mathcal{P} = 1 - \prod_{i=1}^{N_{fs}} \sum_{j=0}^{n_i-1} \frac{e^{-b_i} b_i^j}{j!},$$  \hspace{1cm} (14)

where $n_i$ is the smallest integer satisfying

$$\sum_{j=n_i}^{\infty} \frac{e^{-b_i} b_i^j}{j!} \leq \mathcal{P}_{\text{min}} = \min_i \mathcal{P}_i.$$  \hspace{1cm} (15)

2. Confirmation

An independent confirmation is desirable for any potential discovery, especially for an excess revealed by a data-driven search. Such confirmation may come from an independent experiment, from the same experiment in a different but related final state, from an independent confirmation of the background estimate, or from the same experiment in the same final state using independent data. In the last of these cases, a first sample can be presented to Sleuth to uncover any hints of new physics, and the remaining sample can be subjected to a standard analysis in the region suggested by Sleuth. An excess in this region in the second sample helps to confirm a discrepancy between data and background. If we see hints of new physics in the Run I data, for example, we will be able to predict where new physics might show itself in the upcoming run of the Fermilab Tevatron, Run II.

IV. THE $epX$ DATA SET

As mentioned in Sec. II, we have applied the Sleuth method to DØ data containing one or more electrons and one or more muons. We use a data set corresponding to $108.3 \pm 5.7 \text{ pb}^{-1}$ of integrated luminosity, collected between 1992 and 1996 at the Fermilab Tevatron with the DØ detector. The data set and basic selection criteria are identical to those used in the published $t\bar{t}$ cross section analysis for the dilepton channels [12]. Specifically, we apply global cleanup cuts and select events containing

\footnote{Note that the naive expression $\tilde{\mathcal{P}} = 1 - (1 - \mathcal{P}_{\text{min}})^{N_{fs}}$ is not correct, since this requires $\tilde{\mathcal{P}} \to 1$ for $N_{fs} \to \infty$, and there are indeed an infinite number of final states to examine. The resolution of this paradox hinges on the fact that only an integral number of events can be observed in each final state, and therefore final states with $b_i \ll 1$ contribute very little to the value of $\tilde{\mathcal{P}}$. This is correctly accounted for in the formulation given in Eq. 14.}
The number of expected background events for the populated final states within $e\mu X$. The errors on $e\mu X$ are smaller than on the sum of the individual background contributions obtained from Monte Carlo because of an uncertainty on the number of extra jets arising from initial and final state radiation in the exclusive channels.

| Data set | Fakes | $Z \rightarrow \tau\tau$ | $\gamma^* \rightarrow \tau\tau$ | $WW$ | $t\bar{t}$ | Total |
|----------|-------|----------------|----------------|------|-----|------|
| $e\mu E_T$ | 18.4±1.4 | 25.6±6.5 | 0.5±0.2 | 3.9±1.0 | 0.011±0.003 | 48.5±7.6 |
| $e\mu E_Ttj$ | 8.7±1.0 | 3.0±0.8 | 0.1±0.03 | 1.1±0.3 | 0.4±0.1 | 13.2±1.5 |
| $e\mu E_Tjj$ | 2.7±0.6 | 0.5±0.2 | 0.012±0.006 | 0.18±0.05 | 1.8±0.5 | 5.2±0.8 |
| $e\mu E_Tjjj$ | 0.4±0.2 | 0.07±0.05 | 0.005±0.004 | 0.032±0.009 | 0.7±0.2 | 1.3±0.3 |
| $e\mu X$ | 30.2±1.8 | 29.2±4.5 | 0.7±0.1 | 5.2±0.8 | 3.1±0.5 | 68.3±5.7 |

**TABLE III.** The number of expected background events for the populated final states within $e\mu X$. The errors on $e\mu X$ are smaller than on the sum of the individual background contributions obtained from Monte Carlo because of an uncertainty on the number of extra jets arising from initial and final state radiation in the exclusive channels.

| Final State | Variables |
|------------|-----------|
| $e\mu E_T$ | $p_T^e, \not{E}_T$ |
| $e\mu E_Tj$ | $p_T^e, \not{E}_T, p_T^j$ |
| $e\mu E_Tjj$ | $p_T^e, \not{E}_T, p_T^{j2}$ |
| $e\mu E_Tjjj$ | $p_T^e, \not{E}_T, p_T^{j2} + p_T^{j3}$ |

**TABLE II.** The exclusive final states within $e\mu X$ for which events are seen in the data and the variables used for each of these final states. The variables are selected using the prescription described in Sec. II. Although all final states contain "$e\mu E_T$", no missing transverse energy cut has been applied explicitly; $\not{E}_T$ is inferred from the presence of the muon, following Sec. II.1.

- one or more high $p_T$ ($p_T > 15$ GeV) isolated electrons, and
- one or more high $p_T$ ($p_T > 15$ GeV) isolated muons, with object definitions given in Sec. II.1.

The dominant standard model and instrumental backgrounds to this data set are

- top quark pair production with $t \rightarrow Wb$, and with both $W$ bosons decaying leptonically, one to $e\nu$ (or to $\tau\nu \rightarrow \mu\nu\nu$) and one to $\mu\nu$ (or to $\tau\nu \rightarrow \mu\nu\nu$),
- $W$ boson pair production with both $W$ bosons decaying leptonically, one to $e\nu$ (or to $\tau\nu \rightarrow \mu\nu\nu$) and one to $\mu\nu$ (or to $\tau\nu \rightarrow \mu\nu\nu$),
- $Z/\gamma^* \rightarrow \tau\tau \rightarrow e\mu\nu\nu\nu\nu$, and
- instrumental ("fakes"): $W$ production with the $W$ boson decaying to $\mu\nu$ and a radiated jet or photon being mistaken for an electron, or $b\bar{b}/c\bar{c}$ production with one heavy quark producing an isolated muon and the other a false electron [13].

A sample of $100,000$ $t\bar{t}$ → dilepton events was generated using HERWIG [16], and a $WW$ sample of equal size was generated using PYTHIA [17]. We generated $\gamma^* \rightarrow \tau\tau \rightarrow e\mu\nu\nu\nu$ (Drell-Yan) events using PYTHIA and $Z \rightarrow \tau\tau \rightarrow e\mu\nu\nu\nu$ events using ISAJET [13]. The Drell-Yan cross section is normalized as in Ref. [11]. The cross section for $Z \rightarrow \tau\tau$ is taken to be equal to the published DØ $Z \rightarrow ee$ cross section [2]; the top quark production cross section is taken from Ref. [21]; and the $WW$ cross section is taken from Ref. [22]. The $t\bar{t}$, $WW$, and $Z/\gamma^*$ Monte Carlo events all were processed through GEANT [23] and the DØ reconstruction software. The number and distributions of events containing fake electrons are taken from data, using a sample of events satisfying "bad" electron identification criteria [24].

We break $e\mu X$ into exclusive data sets, and determine which variables to consider in each set using the prescription given in Sec. II. The exclusive final states within $e\mu X$ that are populated with events in the data are listed in Table I. The number of events expected for the various samples and data sets in the populated final states are given in Table II; the number of expected background events in all unpopulated final states in which the number of expected background events is $> 0.001$ are listed in Table IV. The dominant sources of systematic error are given in Table V.

| Final State | Background expected |
|------------|---------------------|
| $e\mu E_Tjjj$ | 0.3 ± 0.15 |
| $e\mu E_T$ | 0.10 ± 0.05 |
| $e\mu$ | 0.04 ± 0.02 |
| $e\mu X\gamma$ | 0.06 ± 0.03 |

**TABLE IV.** The number of expected background events for the unpopulated final states within $e\mu X$. The expected number of events in final states with additional jets is obtained from those listed in the table by dividing by five for each jet. These are all rough estimates, and a large systematic error has been assigned accordingly. Since no events are seen in any of these final states, the background estimates shown here are used solely in the calculation of $\mathcal{P}$ for all $e\mu X$ channels.

V. SENSITIVITY

We choose to consider the $e\mu X$ final state first because it contains backgrounds of mass scale comparable to that expected of the physics responsible for electroweak symmetry breaking. Top quark pair production ($q\bar{q} \rightarrow t\bar{t} \rightarrow W^+W^-b\bar{b}$) and $W$ boson pair production are excellent examples of the type of physics that we would expect the algorithm to find.

Before examining the data, we decided to impose the requirements of AntiCornerSphere and Isolation (see Ap-
result as interesting or more interesting than an empty sample. If one event is observed then there is only one region for Sleuth to consider, and $\mathcal{P}$ is simply the probability for $b \pm \delta b$ to fluctuate up to exactly one event. In Fig. 3(d), for example, the spike at $\mathcal{P} = 1$ contains 62% of the mock experiments, since this is the probability for $0.5 \pm 0.2$ to fluctuate to zero events; the second spike is located at $\mathcal{P} = 0.38$ and contains 28% of the mock experiments, since this is the probability for $0.5 \pm 0.2$ to fluctuate to exactly one event. Similar but less pronounced behavior is seen in Fig. 3(c). Figure 3 shows distributions of $\mathcal{P}$ when the mock samples contain WW and $t\bar{t}$ in addition to the background in Fig. 2. Again, the number of events from each process is allowed to vary within statistical and systematic error. Figure 3 shows that we can indeed find $t\bar{t}$ and/or WW much of the time. Figure 3 shows $\mathcal{P}$ computed for these samples. In over 50% of these samples we find $\mathcal{P}_{[\sigma]}$ to correspond to more than two standard deviations.

**A. Search for WW and $t\bar{t}$ in mock samples**

In this section we provide results from Sleuth for the case in which $Z/\gamma^* \rightarrow \tau\tau$ and fakes are included in the background estimates and the signal from WW and $t\bar{t}$ is “unknown.” We apply the prescription to the exclusive $e\mu X$ final states listed in Table II.

Figure 3 shows distributions of $\mathcal{P}$ for mock samples containing only $Z/\gamma^* \rightarrow \tau\tau$ and fakes, where the mock events are pulled randomly from their parent distributions and the numbers of events are allowed to vary within systematic and statistical errors. The distributions are uniform in the interval $[0, 1]$, as expected, becoming appropriately discretized in the low statistics limit. (When the number of expected background events $b \lesssim 1$, as in Fig. 2(d), it can happen that zero or one events are observed.) If zero events are observed then $\mathcal{P} = 1$, since all hypothetical similar experiments yield a result as interesting or more interesting than an empty sample. If one event is observed then there is only one region for Sleuth to consider, and $\mathcal{P}$ is simply the probability for $b \pm \delta b$ to fluctuate up to exactly one event. In Fig. 3(d), for example, the spike at $\mathcal{P} = 1$ contains 62% of the mock experiments, since this is the probability for $0.5 \pm 0.2$ to fluctuate to zero events; the second spike is located at $\mathcal{P} = 0.38$ and contains 28% of the mock experiments, since this is the probability for $0.5 \pm 0.2$ to fluctuate to exactly one event. Similar but less pronounced behavior is seen in Fig. 3(c). Figure 3 shows distributions of $\mathcal{P}$ when the mock samples contain WW and $t\bar{t}$ in addition to the background in Fig. 2. Again, the number of events from each process is allowed to vary within statistical and systematic error. Figure 3 shows that we can indeed find $t\bar{t}$ and/or WW much of the time. Figure 3 shows $\mathcal{P}$ computed for these samples. In over 50% of these samples we find $\mathcal{P}_{[\sigma]}$ to correspond to more than two standard deviations.

**B. Search for $t\bar{t}$ in mock samples**

In this section we provide results for the case in which $Z/\gamma^* \rightarrow \tau\tau$, fakes, and WW are all included in the back-
samples for these distributions contain \(Z/\gamma\) background includes only \(Z/\gamma\) and fakes. The mock samples for these distributions contain \(WW\) and \(t\bar{t}\) in addition to \(Z/\gamma\) and fakes. The extent to which these distributions peak at small \(P\) can be taken as a measure of Sleuth’s ability to find \(WW\) or \(t\bar{t}\) if we had no knowledge of either final state. The presence of \(WW\) in \(e\mu\) causes the trend toward small values in (a); the presence of \(t\bar{t}\) causes the trend toward small values in (c) and (d); and a combination of \(WW\) and \(t\bar{t}\) causes the signal seen in (b).

FIG. 5. Distributions of \(P\) for the four exclusive final states (a) \(e\mu E_T\), (b) \(e\mu E_T j\), (c) \(e\mu E_T jj\), and (d) \(e\mu E_T jjj\). The background includes only \(Z/\gamma\) \(\rightarrow \tau \tau\) and fakes. The mock samples for these distributions contain \(WW\) and \(t\bar{t}\) in addition to \(Z/\gamma\) \(\rightarrow \tau \tau\) and fakes. The extent to which these distributions peak at small \(P\) can be taken as a measure of Sleuth’s ability to find \(WW\) or \(t\bar{t}\) if we had no knowledge of either final state. The presence of \(WW\) in \(e\mu\) causes the trend toward small values in (a); the presence of \(t\bar{t}\) causes the trend toward small values in (c) and (d); and a combination of \(WW\) and \(t\bar{t}\) causes the signal seen in (b).

Figure 6 shows distributions of \(P\) for mock samples containing \(Z/\gamma\) \(\rightarrow \tau \tau\), \(\tau \tau\) and fakes, and \(WW\), where the mock events are pulled randomly from their parent distributions, and the numbers of events are allowed to vary within systematic and statistical errors. As found in the previous section, the distributions are uniform in the interval \([0, 1]\), becoming appropriately discretized when the expected number of background events becomes \(\lesssim 1\). Figure 6 shows distributions of \(P\) when the mock samples contain \(t\bar{t}\) in addition to \(Z/\gamma\) \(\rightarrow \tau \tau\), \(\tau \tau\) and fakes, and \(WW\). Again, the number of events from each process is allowed to vary within statistical and systematic errors. The distributions in Figs. 6(c) and (d) show that we can indeed find \(t\bar{t}\) much of the time. Figure 6 shows that the distribution of \(P_{[a]}\) is approximately a Gaussian centered at zero of width unity for the case where the background and data both contain \(Z/\gamma\) \(\rightarrow \tau \tau\), \(\tau \tau\) and \(WW\) production, and is peaked in the bin above 2.0 for the same background when the data include \(t\bar{t}\).

FIG. 6. Distribution of \(P_{[a]}\) from combining the four exclusive final states \(e\mu E_T\), \(e\mu E_T j\), \(e\mu E_T jj\), and \(e\mu E_T jjj\). The background includes only \(Z/\gamma\) \(\rightarrow \tau \tau\) and fakes. The mock samples making up the distribution shown as the solid line contain \(WW\) and \(t\bar{t}\) in addition to \(Z/\gamma\) \(\rightarrow \tau \tau\) and fakes, and correspond to Fig. 5. The mock samples making up the distribution shown as the dashed line contain only \(Z/\gamma\) \(\rightarrow \tau \tau\) and fakes, and correspond to Fig. 5. All samples with \(P_{[a]} > 2.0\) appear in the rightmost bin. The fact that \(P_{[a]} > 2.0\) in 50% of the mock samples can be taken as a measure of Sleuth’s sensitivity to finding \(WW\) and \(t\bar{t}\) if we had no knowledge of the existence of the top quark or the possibility of \(W\) boson pair production.

C. New high \(p_T\) physics

We have shown in Secs. V A and V B that the Sleuth prescription and algorithm correctly finds nothing when there is nothing to be found, while exhibiting sensitivity to the expected presence of \(WW\) and \(t\bar{t}\) in the \(e\mu X\) sample. Sleuth’s performance on this “typical” new physics signal is encouraging, and may be taken as some measure of the sensitivity of this method to the great variety of high \(p_T\) physics that it has been designed to find. Making a more general claim regarding Sleuth’s sensitivity to the presence of new physics is difficult, since the sensitivity obviously varies with the characteristics of each candidate theory.

That being said, we can provide a rough estimate of Sleuth’s sensitivity to new high \(p_T\) physics with the following argument. We have seen that we are sensitive to \(WW\) and \(t\bar{t}\) pair production in a data sample corresponding to an integrated luminosity of \(\approx 100\text{ pb}^{-1}\).
These events tend to fall in the region $p_T > 40 \text{ GeV}$, $E_T > 40 \text{ GeV}$, and $\sum p_T > 40 \text{ GeV}$ (if there are any jets at all). The probability that any true $e\mu X$ event produced will make it into the final sample is about 15% due to the absence of complete hermeticity of the DØ detector, inefficiencies in the detection of electrons and muons, and kinematic acceptance. We can therefore state that we are as sensitive to new high $p_T$ physics as we were to the roughly eight $W W$ and $t\bar{t}$ events in our mock samples if the new physics is distributed relative to all standard model backgrounds as $W W$ and $t\bar{t}$ are distributed relative to backgrounds from $Z/\gamma^* \rightarrow \tau\tau$, fakes, and $W W$. The extent to which these distributions peak at small $P$ can be taken as a measure of Sleuth’s sensitivity to finding $t\bar{t}$ if we had no knowledge of the top quark’s existence or characteristics. Note that $P$ is flat in $e\mu E_T$, where the expected number of top quark events is negligible, peaks slightly toward small values in $e\mu E_T j$, and shows a marked low peak in $e\mu E_T jj$ and $e\mu E_T jjj$.

VI. RESULTS

In the previous section we studied what can be expected when Sleuth is applied to $e\mu X$ mock samples. In this section we confront Sleuth with data. We observe 39 events in the $e\mu E_T$ final state, 13 events in $e\mu E_T j$, 5 events in $e\mu E_T jj$, and a single event in $e\mu E_T jjj$, in good agreement with the expected background in Table III. We proceed by first removing both $W W$ and $t\bar{t}$ from the background estimates, and next by removing only $t\bar{t}$, to search for evidence of these processes in the data. Finally, we include all standard model processes in the background estimates and search for evidence of new physics.

A. Search for $WW$ and $t\bar{t}$ in data

The results of applying Sleuth to DØ data with only $Z/\gamma^* \rightarrow \tau\tau$, fakes, and $W W$ in the background estimate are shown in Table V and Fig. 10. Sleuth finds indications of an excess in the $e\mu E_T$ and $e\mu E_T jj$ states, presumably reflecting the presence of $W W$ and $t\bar{t}$, respectively. The results for the $e\mu E_T j$ and $e\mu E_T jjj$ final states are consistent with the results in Fig. 5. Defining $r'$ as the distance of the data point from $(0,0,0)$ in the unit box...
Our analysis of the background distribution is shown in Table VII and Fig. 11. Sleuth finds an indication of excess in the $e\mu E_T$, $e\mu E_Tj$, and $e\mu E_Tjj$ channels when WW and $t\bar{t}$ are not included in the background. Sleuth identifies a region of excess in the $e\mu E_T$ and $e\mu E_Tjj$ final states, presumably indicating the presence of WW and $t\bar{t}$ in the data. In units of standard deviation, $\mathcal{P}_{[a]} = 1.9$.

### Table VI. Summary of results on the $e\mu E_T$, $e\mu E_Tj$, $e\mu E_Tjj$, and $e\mu E_Tjjj$ channels when WW and $t\bar{t}$ are not included in the background.

| Data set   | \(\mathcal{P}\) |
|------------|------------------|
| $e\mu E_T$ | 0.008            |
| $e\mu E_Tj$ | 0.34             |
| $e\mu E_Tjj$ | 0.01             |
| $e\mu E_Tjjj$ | 0.38             |
| $\mathcal{P}$ | 0.03             |

The results of applying Sleuth to the data with $Z/\gamma^* \rightarrow \tau\tau$, fakes, and WW included in the background estimate are shown in Table VII and Fig. 11. Sleuth finds an indication of excess in the $e\mu E_Tjjj$ events, presumably indicating the presence of $t\bar{t}$. The results for the $e\mu E_T$, $e\mu E_Tj$, and $e\mu E_Tjjj$ final states are consistent with the results in Fig. 8. The $t\bar{t}$ candidates from DO’s recent analysis are the three events with largest $r'$ in the $e\mu E_Tjjj$ sample, and the single event in the $e\mu E_Tjjj$ sample, shown in Fig. 9.

### B. Search for $t\bar{t}$ in data

The results of applying Sleuth to the data with $Z/\gamma^* \rightarrow \tau\tau$, fakes, and WW included in the background are shown in Table VII and Fig. 11. Sleuth finds an indication of excess in the $e\mu E_Tjjj$ events, presumably indicating the presence of $t\bar{t}$. The results for the $e\mu E_T$, $e\mu E_Tj$, and $e\mu E_Tjjj$ final states are consistent with the results in Fig. 8. The $t\bar{t}$ candidates from DO’s recent analysis are the three events with largest $r'$ in the $e\mu E_Tjjj$ sample, and the single event in the $e\mu E_Tjjj$ sample, shown in Fig. 11.

A comparison of this result with one obtained using a dedicated top quark search illustrates an important difference between Sleuth’s result and the result from a dedicated search. DO announced its discovery of the top
sample, a 1.9σ “effect,” when complete ignorance of the top quark is feigned. When we take into account the fact that we have also searched in all of the final states listed in Table II, we find \( \bar{P} = 0.11 \), a 1.2σ “effect.” The difference between the 2.75σ “effect” seen with a dedicated top quark search and the 1.2σ “effect” that Sleuth reports in \( e\mu X \) lies partially in the fact that Sleuth is not optimized for \( tt \); and partially in the careful accounting of the many new physics signatures that Sleuth considered in addition to \( tt \) production, and the correspondingly many new physics signals that Sleuth might have discovered.

C. Search for physics beyond the standard model

In this section we present Sleuth’s results for the case in which all standard model and instrumental backgrounds are considered in the background estimate: \( Z/\gamma^* \rightarrow \tau\tau, \) fakes, \( WW, \) and \( tt \). The results are shown in Table VIII and Fig. 12. We observe excellent agreement with the standard model. We conclude that these data contain no evidence of new physics at high \( p_T \), and calculate that a fraction \( \bar{P} = 0.72 \) of hypothetical similar experimental runs would produce a more significant excess than any observed in these data. Recall that we are sensitive to new high \( p_T \) physics with production cross section \( \times \) branching ratio into this final state as described in Sec. V C.

TABLE VIII. Summary of results on all final states within \( e\mu X \) when all standard model backgrounds are included. The unpopulated final states (listed in Table IV) have \( P = 1.0 \); these final states are included in the calculation of \( \bar{P} \). We observe no evidence for the presence of new high \( p_T \) physics.

| Data set   | \( P \)  |
|------------|---------|
| \( e\mu E_T \) | 0.14    |
| \( e\mu E_T jj \) | 0.45    |
| \( e\mu E_T \) | 0.31    |
| \( e\mu E_T jj jj \) | 0.71    |
| \( \bar{P} \) | 0.72    |

VII. CONCLUSIONS

We have developed a quasi-model-independent technique for searching for the physics responsible for stabilizing electroweak symmetry breaking. Our prescription involves the definition of final states and the construction of a rule that identifies a set of relevant variables for any particular final state. An algorithm (Sleuth) systematically searches for regions of excess in those variables, and quantifies the significance of any observed excess. This technique is sufficiently \textit{a priori} that it allows an \textit{ex post facto}, quantitative measure of the degree to which curious

| Data set   | \( P \)  |
|------------|---------|
| \( e\mu E_T \) | 0.16    |
| \( e\mu E_T jj \) | 0.45    |
| \( e\mu E_T jj jj \) | 0.03    |
| \( e\mu E_T jj jj jj \) | 0.41    |
| \( \bar{P} \) | 0.11    |

TABLE VII. Summary of results on the \( e\mu E_T \), \( e\mu E_T jj \), \( e\mu E_T jj jj \), and \( e\mu E_T jj jj jj \) channels when \( tt \) production is not included in the background. Sleuth identifies a region of excess in the space of variables in Table I to a uniform distribution in the unit box. The darkened points define the region Sleuth found most interesting. The interesting region in the \( e\mu E_T jj \) sample presumably indicates the presence of \( tt \) in the data. In units of standard deviation, \( \bar{P} = 1.2 \).

FIG. 11. Positions of data points following the transformation of the background from the three sources \( Z/\gamma^* \rightarrow \tau\tau, \) fakes, and \( WW \) in the space of variables in Table I to a uniform distribution in the unit box. The darkened points define the region Sleuth found most interesting. The interesting region in the \( e\mu E_T jj \) sample presumably indicates the presence of \( tt \). We find \( \bar{P} = 0.11 \) (\( \bar{P} = 1.2 \)).

C. Search for physics beyond the standard model

In this section we present Sleuth’s results for the case in which all standard model and instrumental backgrounds are considered in the background estimate: \( Z/\gamma^* \rightarrow \tau\tau, \) fakes, \( WW, \) and \( tt \). The results are shown in Table VIII and Fig. 12. We observe excellent agreement with the standard model. We conclude that these data contain no evidence of new physics at high \( p_T \), and calculate that a fraction \( \bar{P} = 0.72 \) of hypothetical similar experimental runs would produce a more significant excess than any observed in these data. Recall that we are sensitive to new high \( p_T \) physics with production cross section \( \times \) branching ratio into this final state as described in Sec. V C.

TABLE VIII. Summary of results on all final states within \( e\mu X \) when all standard model backgrounds are included. The unpopulated final states (listed in Table IV) have \( P = 1.0 \); these final states are included in the calculation of \( \bar{P} \). We observe no evidence for the presence of new high \( p_T \) physics.

| Data set   | \( P \)  |
|------------|---------|
| \( e\mu E_T \) | 0.14    |
| \( e\mu E_T jj \) | 0.45    |
| \( e\mu E_T jj jj \) | 0.31    |
| \( e\mu E_T jj jj jj \) | 0.71    |
| \( \bar{P} \) | 0.72    |
FIG. 12. Positions of the data points following the transformation of the background from $Z/\gamma^* \rightarrow \tau \tau$, fakes, $WW$, and $t\bar{t}$ sources in the space of variables in Table I to a uniform distribution in the unit box. The darkened points define the region that Sleuth chose. We find $\tilde{\mathcal{P}} = 0.72$, and distributions that are all roughly uniform and consistent with background. No evidence for new high $p_T$ physics is observed.

events are interesting. After demonstrating the sensitivity of the method, we have applied it to the set of events in the semi-inclusive channel $e\mu X$. Removing $WW$ and $t\bar{t}$ from the calculated background, we find indications of these signals in the data. Including these background channels, we find that these data contain no evidence of new physics at high $p_T$. A fraction $\mathcal{P} = 0.72$ of hypothetical similar experimental runs would produce a more significant excess than any observed in these data.

ACKNOWLEDGMENTS

We thank the staffs at Fermilab and at collaborating institutions for contributions to this work, and acknowledge support from the Department of Energy and National Science Foundation (USA), Commissariat à L’Énergie Atomique and CNRS/Institut National de Physique Nucléaire et de Physique des Particules (France), Ministry for Science and Technology and Ministry for Atomic Energy (Russia), CAPES and CNPq (Brazil), Departments of Atomic Energy and Science and Education (India), Colciencias (Colombia), CONACyT (Mexico), Ministry of Education and KOSEF (Korea), CONICET and UBACyT (Argentina), A.P. Sloan Foun-
$b(x) = \frac{1}{5}e^{-x/5}$ and we let $y = 1 - e^{-x/5}$; then $y$ is distributed uniformly between 0 and 1. The situation is more complicated when the background is given to us as a set of Monte Carlo points that cannot be described by a simple parameterization, and it is further complicated when these points live in several dimensions.

There is a unique solution to this problem in one dimension, but an infinity of solutions in two or more dimensions. Not all of these solutions are equally reasonable, however — there are two additional properties that the solution should have.

- Axes should map to axes. If the data live in a three-dimensional space in the octant with all coordinates positive, for example, then it is natural to map the coordinate axes to the axes of the box.
- Points that are near each other should map to points that are near each other, subject to the constraint that the resulting background probability distribution be flat within the unit box.

This somewhat abstract and not entirely well-posed problem is helped by considering an analogous physical problem:

The height of the sand in a $d$-dimensional unit sandbox is given by the function $b(\vec{x})$, where $\vec{x}$ is a $d$-component vector. (The counting of dimensions is such that a physical sandbox has $d = 2$.) We take the $d$-dimensional lid of the sandbox and squash the sand flat. The result of this squashing is that a sand grain at position $\vec{x}$ has moved to a new position $\vec{y}$, and the new function $b'(\vec{y})$ describing the height of the sand is a constant. Given the function $b(\vec{x})$, determine the mapping $\vec{x} \rightarrow \vec{y}$.

For this analogy to help, the background first needs to be put “in the sandbox.” Each of the background events must also have the same weight (the reason for this to be put “in the sandbox.” Each of the background events and map each of them to a point on a uniform grid within the box. The previous paragraph defines a mapping from the original variables into the unit sandbox; this step defines a mapping from a lumpy distribution in the sandbox to a flat distribution. The mapping is continued to the entire space by interpolating between the sampled background events.

The mapping to the grid is done by first assigning each sampled background point to an arbitrary grid point. Each background point $i$ is some distance $d_{ij}$ away from the grid point $j$ with which it is paired. We then loop over pairs of background points $i$ and $i'$, which are associated with grid points $j$ and $j'$, and swap the associations (associate $i$ with $j'$ and $i'$ with $j$) if $\max(d_{ij}, d_{i'j'}) > \max(d_{i'j}, d_{ij'})$. This looping and swapping is continued until an equilibrium state is reached.

APPENDIX C: REGION CRITERIA

In Sec. \[\text{III B 3}\] we introduced the formal notion of region criteria — properties that we require a region to have for it to be considered by Sleuth. The two criteria that we have decided to impose in the analysis of the $\epsilon\mu X$ data are Isolation and AntiCornerSphere.

a. Isolation We want the region to include events that are very close to it. We define $\xi = \frac{1}{4}N_{\text{data}}^{-\frac{1}{2}}$, as a measure of the mean distance between data points in their transformed coordinates, and call a region isolated if there exist no data points outside the region that are closer than $\xi$ to a data point inside the region. We generalize this boolean criterion to the interval $[0, 1]$ by defining

$$e_{R}^{\text{Isolation}} = \min \left( 1, \frac{\min |(\vec{x})_{\text{in}} - (\vec{x})_{\text{out}}|}{2\xi} \right) ,$$

where the minimum is taken over all pairwise combinations of data points with $(\vec{x})_{\text{in}}$ inside $R$ and $(\vec{x})_{\text{out}}$ outside $R$.

b. AntiCornerSphere One must be able to draw a sphere centered on the origin of the unit box containing all data events outside the region and no data events inside the region. This is useful if the signal is expected to lie in the upper right-hand corner of the unit box. We generalize this boolean criterion to the interval $[0, 1]$ as described in Sec. \[\text{III B 3}\].

A number of other potentially useful region criteria may be imagined. Among those that we have considered are Connectivity, Convexity, Peg, and Hyperplanes. Although we present only the boolean forms of these criteria here, they may be generalized to the interval $[0, 1]$ by introducing the scale $\xi$ in the same spirit as above.

c. Connectivity We generally expect a discovery region to be one connected subspace in the variables we use, rather than several disconnected subspaces. Although
one can posit cases in which the signal region is not con-  

nected (perhaps signal appears in the two regions \( \eta > 2 \)  

and \( \eta < -2 \)), one should be able to easily avoid this with  

an appropriate choice of variables. (In this example, we  

should use \(| \eta | \) rather than \( \eta \).) We defined the concept  

of neighboring data points in the discussion of regions in  

Sec. III B 2. A connected region is defined to be a region  

in which given any two points \( p_a \) and \( p_b \) of \( p \)  

there exists a list of points \( p_1 = a, p_2, \ldots, p_{n-1}, p_n = b \)  

such that all the \( p_i \) are in the region and \( p_{i+1} \) is a neigh-  

bor of \( p_i \).

d. Convexity We define a non-convex region as a re-  

gion defined by a set of \( N \) data points \( P \), such that there  

exists a data point \( \vec{p} \) not within \( P \) satisfying  

\[
\sum_{i=1}^{N} \vec{p}_i \lambda_i = \vec{p} \quad \text{(C2)} 
\]

\[
\sum_i \lambda_i = 1 \quad \text{(C3)} 
\]

\[
\lambda_i \geq 0 \quad \forall \ i, \quad \text{(C4)} 
\]

for suitably chosen \( \lambda_i \), where \( \vec{p}_i \) are the points within \( P \).  

A convex region is then any region that is not non-convex;  

intuitively, a convex region is one that is “roundish,”  

without protrusions or intrusions.

e. Peg We may want to consider only regions that  

live on the high tails of a distribution. More generally,  

we may want to only consider regions that contain one  

or more of \( n \) specific points in variable space. Call this  

set of points \( \bar{x}_i \), where \( i = 1, \ldots, n \). We transform these  

points exactly as we transformed the data in Sec. III B  

to obtain a set of points \( \bar{y}_i \) that live in the unit box. A  

region \( R \) is said to be pegged to these points if there exists  

at least one \( i \in 1, \ldots, n \) such that the closest data point  
to \( \bar{y}_i \) lies within \( R \).

f. Hyperplanes Connectivity and Convexity are cri-  

teria that require the region to be “reasonably-shaped,”  

while Peg is designed to ensure that the region is “in a be-  

lievable location.” It is possible, and may at times be de-  

sirable, to impose a criterion that judges both shape and  
location simultaneously. A region \( R \) in a \( d \)-dimensional  
unit box is said to satisfy Hyperplanes if, for each data  
point \( p \) inside \( R \), one can draw a \((d - 1)\)-dimensional  
hyperplane through \( p \) such that all data points on the  
side of the hyperplane containing the point \( \vec{1} \) (the “up- 

per right-hand corner of the unit box”) are inside \( R \).

More complicated region criteria may be built from  
combinations and variations of these and other basic ele-  
ments.

APPENDIX D: SEARCH HEURISTIC DETAILS

The heuristic Sleuth uses to search for the region of  
greatest excess may usefully be visualized as a set of rules  
for an amoeba to move within the unit box. We monitor  
the amoeba’s progress by maintaining a list of the most  
interesting region of size \( N \) (one for each \( N \)) that the  
amoeba has visited so far. At each state, the amoeba is  
the region under consideration, and the rules tell us what  
region to consider next.

The initial location and size of the amoeba is deter-  
mined by the following rules for seeding:

1. If we have not yet searched this data set at all, the  
starting amoeba fills the entire box.

2. Otherwise, the amoeba starts out as the region  
around a single random point that has not yet in- 
habited a “small” region that we have considered  
so far. We consider a region \( R \) to be small if adding  
or removing an individual point can have a sizeable  
effect on the \( p_R^N \); in practice, a region is small if \( N \lesssim 20 \).

3. If there is no point that has not yet inhabited a  
small region that we have considered so far, the  
search is complete.

At each stage, the amoeba either grows or shrinks. It  

grows by attempting to grow. The rules for growth are:

1. Allow the amoeba to encompass a neighboring data  
point. Force it to encompass any other data points  
necessary to make the expanded amoeba satisfy all  
criteria. Check to see whether the \( p_R^{N+1} \) of the ex- 

panded amoeba is less than the \( p_R^N \) of the region  
on the list of the same size. If so, the amoeba has  
successfully grown, the list of the most interesting  
regions is updated, and the amoeba tries to grow  
again. If not, the amoeba shrinks back to its former  
size and repeats the same process using a different  
nearby data point.

2. If the amoeba has tried all neighboring data points  
and has not successfully grown, it shrinks.

The rules for shrinking are:

1. Force the amoeba to relinquish the data point that  
owns the most background, subject to the require- 
ment that the resulting shrunken amoeba be con-  
sistent with the criteria.

2. If the amoeba has shrunk out of existence or can  
shrink no further, we kill this amoeba and reseed.

The result of this process is a list of regions of length  
\( N_{\text{data}} \) (one region for each \( N \)), such that the \( N^{th} \) region  
in the list is the most interesting region of size \( N \) found  
in the data set.
[1] L. Hall, J. Lykken and S. Weinberg, Phys. Rev. D 27, 2359 (1983).
[2] J. Gunion, H. E. Haber, G. L. Kane, and S. Dawson, The Higgs Hunter’s Guide (Addison-Wesley, Redwood City, 1990).
[3] M. Hohlmann, Ph. D. thesis, University of Chicago (1997).
[4] R. M. Barnett and L. J. Hall, Phys. Rev. Lett. 77, 3506 (1996).
[5] P. Nath and R. Arnowitt, Mod. Phys. Lett. A 2, 331 (1987); R. Barbieri et al., Nucl. Phys. B 367, 28 (1991); H. Baer and X. Tata, Phys. Rev. D 47, 2739 (1993); J. Lopez et al., Phys. Rev. D 48, 2062 (1993).
[6] G. Valencia and S. Willenbrock, Phys. Rev. D 50, 6843 (1994).
[7] H. Haber and G. Kane, Phys. Rev. 117, 75 (1985).
[8] S. Weinberg, Phys. Rev. D 13, 974 (1976); ibid. 19, 1277 (1979); L. Susskind, Phys. Rev. D 20, 2619 (1979); S. Dimopoulos and L. Susskind, Nuc. Phys. B 155, 237 (1979); E. Eichten and K. Lane, Phys. Lett. B 90, 125 (1980).
[9] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali, Phys. Lett. B 429, 263 (1998).
[10] DØ Collaboration, S. Abachi et al., Phys. Rev. Lett. 79, 1197 (1997).
[11] DØ Collaboration, S. Abachi et al., Nucl. Instr. and Methods A338, 185 (1994).
[12] DØ Collaboration, S. Abachi et al., Phys. Rev. Lett. 79, 1203 (1997).
[13] DØ Collaboration, S. Abachi et al., Phys. Rev. D 52, 4877 (1995).
[14] DØ Collaboration, S. Abachi et al., Phys. Rev. Lett. 78, 2070 (1997).
[15] K. Clarkson, http://cm.bell-labs.com/netlib/voronoi/hull.html (1996).
[16] G. Marchesini et al., hep-ph/9607393, 1996 (unpublished); G. Marchesini et al., Comp. Phys. Comm. 67, 465 (1992). We used v5.7.
[17] T. Sjöstrand, Comp. Phys. Comm. 82, 74 (1994). We used v5.7.
[18] F. Paige and S. Protopopescu, BNL Report No. 38304 1986 (unpublished). We used v7.22 with CTEQ2L parton distribution functions.
[19] J. McKinley, Ph.D. thesis, Michigan State University, 1996 (unpublished).
[20] DØ Collaboration, B. Abbott et al., Phys. Rev. D 61, 072001 (2000).
[21] E. Laenen, J. Smith and W. van Neerven, Phys. Lett. B 321, 254 (1994).
[22] J. Ohnemus, Phys. Rev. D 44, 1403 (1991).
[23] R. Brun and F. Carminati, CERN Program Library Long Writeup W5013, 1993 (unpublished).
[24] DØ Collaboration, S. Abachi et al., Phys. Rev. D 58, 052001 (1998).
[25] H. Singh, Ph.D. thesis, University of California at Riverside, 1999 (unpublished).
[26] DØ Collaboration, S. Abachi et al., Phys. Rev. Lett. 74, 2632 (1995).
[27] L. Holmström, S. R. Sain, and H. E. Miettinen, Comp. Phys. Commun. 88, 195 (1995).