Testing Theories in Particle Physics Using Maximum Likelihood and Adaptive Bin Allocation

Bruce Knuteson[1] and Ricardo Vilalta[2]

[1] Laboratory for Nuclear Science, Massachusetts Institute of Technology
77 Massachusetts Ave. Cambridge, MA 02139-4307, USA
knuteson@mit.edu
[2] Department of Computer Science, University of Houston
4800 Calhoun Rd., Houston TX 77204-3010, USA
vilalta@cs.uh.edu

Abstract. We describe a methodology to assist scientists in quantifying the degree of evidence in favor of a new proposed theory compared to a standard baseline theory. The figure of merit is the log-likelihood ratio of the data given each theory. The novelty of the proposed mechanism lies in the likelihood estimations; the central idea is to adaptively allocate histogram bins that emphasize regions in the variable space where there is a clear difference in the predictions made by the two theories. We describe a software system that computes this figure of merit in the context of particle physics, and describe two examples conducted at the Tevatron Ring at the Fermi National Accelerator Laboratory. Results show how two proposed theories compare to the Standard Model and how the likelihood ratio varies as a function of a physical parameter (e.g., by varying the particle mass).

1 Introduction

Common to many scientific fields is the problem of comparing two or more competing theories based on a set of actual observations. In particle physics, for example, the behavior of Nature at small distance scales is currently well described by the Standard Model. But compelling arguments suggest the presence of new phenomena at distance scales now being experimentally probed, and there exists a long array of proposed extensions to the Standard Model.

The problem of assessing theories against observations can be solved in various ways. Some previous work bearing an artificial intelligence flavor has attempted to use observations to explain processes in both particle physics and astrophysics [4]. From a statistical view, a common solution is to use a maximum-likelihood approach [1, 2], that selects the theory T maximizing $\mathrm{P}(\mathcal{D}|T)$ (i.e., the conditional probability of a set of actual observations $\mathcal{D}$ assuming T is true). Implicit to this methodology is the–often false–assumption that the form of the distributions characterizing the set of competing theories is known. In practice, a scientist suggests a new theory in the form of new equations or new parameters (e.g., new suggested mass for an elementary particle). In particle physics, a software is then used to simulate the response of the particle detector if the new proposed theory T were true, resulting in a data file made of Monte Carlo events from which one can estimate the true distribution characterizing T. At that point one can compare how close T matches the actual observations (stored in $\mathcal{D}$) obtained from real particle colliders.

To estimate the true distribution of a theory T, we take the Monte Carlo data and follow a histogram approach [5]. We create a series of bins $\{b_k\}$ over the variable space and attempt to predict the number of events expected in every bin b_k if theory T were true. The novelty of our approach lies in the adaptive mechanism behind this bin allocation. Bins are selected to emphasize regions where the number of events predicted by T is significantly different from those predictions generated by competing theories, in a sense discovering regions in the variable space where a discrepancy among theories is evident.

This paper is organized as follows. Section 2 provides background information and notation. Section 3 provides a general description of the mechanism to compute likelihood ratios. Section 4 describes a solution to the problem of adaptive bin allocation. Section 5 reports on the experimental analysis. Lastly, Section 6 gives a summary and discusses future work.

2 Background Information and Notation

In modern particle accelerators, collisions of particles travelling at nearly the speed of light produce debris that is captured by signals from roughly one million channels of readout electronics. We call each collision an *event*. Substantial processing of the recorded signals leads to an identification of the different objects (e.g., electrons ($e^{\pm}$), muons ($\mu^{\pm}$), taus ($\tau^{\pm}$), photons (γ), jets (j), b-jets (b), neutrinos (ν), etc.) that have produced any particular cluster of energy in the detector. Each object is characterized by roughly three variables, corresponding to the three components of the particle's momentum. An event is represented as the composition of many objects, one for each object detected out of the collision. These kinematic variables can be usefully thought of as forming a *variable space*.

We store events recorded from real particle accelerators in a dataset $\mathcal{D} = \{\mathbf{e}_i\}$, where each event $\mathbf{e} = (a_1, a_2, \cdots, a_n) \in A_1 \times A_2 \times \cdots \times A_n$ is a variable vector characterizing the objects identified on a particular collision. We assume numeric variables only (i.e., $a_i \in \Re$) and that $\mathcal{D}$ consists of independently and identically distributed (i.i.d.) events obtained according to a fixed but unknown joint probability distribution in the variable space.

We assume two additional datasets, $\tilde{D_n}$ and $\tilde{D_s}$, made of discrete Monte Carlo events generated by a detector simulator designed to imitate the behavior of a real particle collider. The first dataset assumes the realization of a new proposed theory T_N; the second dataset is generated under the assumption that the Standard Model T_S is true. Events follow the same representation on all three datasets.

3 Overview of Main Algorithm

In this section we provide a general description of our technique. To begin, assume a physicist puts forth an extension to the Standard Model through a new theory T_N. We define our metric of interest as follows:

$$\mathcal{L}(T_N) = \log_{10} \frac{\mathrm{P}(\mathcal{D}|T_N)}{\mathrm{P}(\mathcal{D}|T_S)} \tag{1}$$

where $\mathcal{D}$ is the set of actual observations obtained from real particle colliders. Metric $\mathcal{L}$ can be conveniently thought of as units of evidence for or against theory T_N. The main challenge behind the computation of $\mathcal{L}$ lies in estimating the likelihoods $\mathrm{P}(\mathcal{D}|\cdot)$. We explain each step next.

3.1 Partitioning Events into Final States

Each event (i.e., each particle collision) may result in the production of different objects, and thus it is appropriate to represent events differently. As an example, one class of events may result in the production of an electron; other events may result in the production of a muon. The first step consists of partitioning the set of events into subsets, where each subset comprises events that produced the same type of objects. This partitioning is orthogonal; each event is placed in one and only one subset, also called *final state*. Let m be the number of final states; the partitioning is done on all three datasets: $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{m}$, $\tilde{D_n} = \{\tilde{D_{ni}}\}_{i=1}^{m}$, and $\tilde{D_s} = \{\tilde{D_{si}}\}_{i=1}^{m}$. Each particular set of subsets $\{\mathcal{D}_i, \tilde{D_{ni}}, \tilde{D_{si}}\}$ is represented using the same set of variables. Estimations obtained from each set of subsets are later combined into a single figure (Section 3.3).

3.2 Computation of Binned Likelihoods

The second step consists of estimating the likelihoods $\mathrm{P}(\mathcal{D}|\cdot)$ adaptively by discovering regions in the variable space where there is a clear difference in the number of Monte Carlo event predictions made by T_N and T_S. Since we treat each subset of events (i.e., each final state) independently (Section 3.1), in this section we assume all calculations refer to a single final state (i.e. a single set of subsets of events $\{\mathcal{D}_i, \tilde{D_{ni}}, \tilde{D_{si}}\}$).

We begin by putting aside for a moment the real-collision dataset $\mathcal{D}_i$. The discrete Monte Carlo events predicted by T_N and T_S in datasets $\tilde{D_{ni}}$ and $\tilde{D_{si}}$ are used to construct smooth probability density estimates $\mathrm{P}_i(\mathbf{e}|T_N)$ and $\mathrm{P}_i(\mathbf{e}|T_S)$. Each density estimate assumes a mixture of Gaussian models:

$$\mathrm{P}_i(\mathbf{e}|T) = \mathrm{P}_i^T(\mathbf{e}) = \sum_{l=1}^{r} \alpha_l \; \phi(\mathbf{e}; \mu_l, \varSigma_l) \tag{2}$$

where r is the number of Gaussian models used to characterize the theory T under consideration. The mixing proportions α_l are such that $\sum_l \alpha_l = 1$, and $\phi(\cdot)$ is a multivariate normal density function:

$$\phi(\mathbf{e}; \mu, \varSigma) = \frac{1}{(2\pi)^{d/2}|\varSigma|^{1/2}} \; \exp\,[-\frac{1}{2}(\mathbf{e} - \mu)^t \varSigma^{-1}(\mathbf{x} - \mu)] \tag{3}$$

where $\mathbf{e}$ and μ are d-component vectors, and $|\varSigma|$ and $\varSigma^{-1}$ are the determinant and inverse of the covariance matrix.

At this point we could follow the traditional approach to Maximum Likelihood estimation by using the real-collision dataset $\mathcal{D}_i$ and the above probability density estimates:

$$\mathrm{P}(\mathcal{D}_i|T) = \prod_j \mathrm{P}_i(\mathbf{e}_j|T) = \prod_j \mathrm{P}_i^T(\mathbf{e}_j) \tag{4}$$

where T takes on T_N or T_S and the index j goes along the events in $\mathcal{D}_i$.

The densities $\mathrm{P}_i(\mathbf{e}|T)$ can in principle be used to compute an unbinned likelihood ratio. But in practice, this ratio can suffer from systematic dependence on the details of the smoothing procedure. Over-smoothed densities cause a bias in favor of distributions with narrow Gaussians, while the use of under-smoothed densities cause undesired dependence on small data irregularities. The calculation of a binned likelihood ratio in the resulting discriminant reduces the dependence on the smoothing procedure, and has the additional

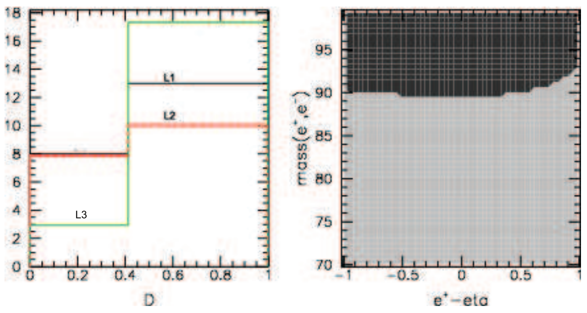

Fig. 1. (Left) The optimally-binned histogram of the discriminant D for the predictions of T_S (line L2), T_N (line L3), and real data $\mathcal{D}$ (line L1). (Right) The mapping of the bins in D back into regions in the original variable space. The dark region corresponds to points $\mathbf{e}$ in the variable space for which $\mathrm{D}(\mathbf{e}) < \theta$; the light region corresponds to points $\mathbf{e}$ in the variable space for which $\mathrm{D}(\mathbf{e}) > \theta$ ($\theta = 0.4$).

advantage that it can be used directly to highlight regions in the variable space where predictions from the two competing theories T_N and T_S differ significantly. We thus propose to follow a histogram technique [5] as follows.

Constructing a Binned Histogram

We begin by defining the following discriminant function:

$$\mathrm{D}(\mathbf{e}) = \frac{\mathrm{P}_i(\mathbf{e}|T_N)}{\mathrm{P}_i(\mathbf{e}|T_N) + \mathrm{P}_i(\mathbf{e}|T_S)} \tag{5}$$

The discriminant function D takes on values between zero and unity, approaching zero in regions in which the number of events predicted by the Standard Model T_S greatly exceeds the number of events predicted by the new proposed theory T_N, and approaching unity in regions in which the number of events predicted by T_N greatly exceeds the number of events predicted by T_S. We employ function D for efficiency reasons: it captures how the predictions of T_N and T_S vary in a single dimension.

We use D to adaptively construct a binned histogram. We compute the value of the discriminant D at the position of each Monte Carlo event predicted by T_N (i.e., every event contained in $\tilde{D_n}$) and T_S (i.e. every event contained in $\tilde{D_s}$). The resulting distributions in D are then divided into a set of bins that maximize an optimization function. This is where our adaptive bin allocation strategy technique is invoked (explained in detail in Section 4). The result is a set of bins that best differentiate the predictions made by T_N and T_S. The output of the Adaptive-Bin-Allocation algorithm is an estimation of the conditional probability $\mathrm{P}(\mathcal{D}_i|T)$.

As an illustration, Figure 1 (left) shows the resulting binned histogram in D for a real scenario with a final state e^+e^- (i.e., electron and positron). The Adaptive-Bin-Allocation algorithm chooses to consider only two bins, placing a bin edge at $\mathrm{D} = 0.4$. Note events from T_S (line L2) tend to lie at values for which $\mathrm{D}(\mathbf{e})$ is small, and events from T_N (line L3) tend to lie at values for which $\mathrm{D}(\mathbf{e})$ is large.

Figure 1 (right) shows how the two bins in the discriminant map back onto the original variable space defined on $m_{e^+e^-}$ (the invariant mass of the electron positron pair), and positron pseudorapidity. The dark region corresponds to points $\mathbf{e}$ in the variable space for which $\mathrm{D}(\mathbf{e}) < 0.4$; similarly the light region corresponds to points $\mathbf{e}$ for which $\mathrm{D}(\mathbf{e}) > 0.4$. Each region is assigned a binned probability (Section 4); all probabilities are then combined into a final state probability $\mathrm{P}(\mathcal{D}_i|T)$.

3.3 Combining Likelihoods and Incorporating Systematic Errors

Once we come up with an estimation of $\mathrm{P}(\mathcal{D}_i|T)$, the next step consists of combining all probabilities from individual final states into a single probability for the entire experiment through the product $\mathrm{P}(\mathcal{D}|T) = \prod_i \mathrm{P}(\mathcal{D}_i|T)$, where T takes on T_N or T_S and the index i goes along all final states. As a side note, a single particle accelerator has normally several experiments running that can also be combined through such products.

Finally, systematic uncertainties are introduced into the analysis to reflect possible imperfections in the modelling of the response of the physical detector. There are usually roughly one dozen sources of systematic error, ranging from possible systematic bias in the measurements of particle energies to an uncertainty in the total amount of data collected.

4 Adaptive Bin Allocation

We now explain in detail our approach to estimate the likelihood $\mathrm{P}(\mathcal{D}_i|T)$ for a particular final state. To begin, assume we have already computed the value of the discriminant D at the position of each Monte Carlo event predicted by T_N and T_S (Section 3.2), and decided on a particular form of binning that partitions D into a set of bins $\{b_k\}$. Let $\mu_{k|T}$ be the number of events expected in bin k if theory T is true[3]. Often in the physical sciences the distribution of counts in each bin is Poisson; this is assumed in what follows. The probability of observing λ_k events in a particular bin k is defined as:

$$\mathrm{P}(\lambda_k|T) = \frac{e^{-\mu_{k|T}} {\mu_{k|T}}^{\lambda_k}}{\lambda_k!} \tag{6}$$

Now, the probability of observing the real data D_i assuming the correctness of T and neglecting correlated uncertainties among the predictions of T in each bin, is simply:

$$\mathrm{P}(D_i|T) = \prod_k \mathrm{P}(\lambda_k|T) \tag{7}$$

where the index k runs along the bins and λ_k is the number of events observed in the real data D_i within bin k.

The question we now pose is how should the bins be chosen? Many finely spaced bins allow finer sampling of differences between T_N and T_S, but introduce a larger uncertainty in the prediction within each bin (i.e., the difference in the events predicted by T_N and T_S under finely spaced bin comes with low confidence levels). On the other hand, a few coarsely spaced bins allow only coarse sampling of the distributions predicted by T_N and T_S, but the predictions within each bin are more robust. The question at hand is not only how many bins to use, but also where to place their edges along the discriminant D [3].

4.1 Searching the Space of Binnings

In selecting an optimal binning we focus our analysis on the two theories T_N and T_S exclusively (choosing a set of optimal bins is independent of the real data used for theory validation). Our goal is to produce a set of bins $\{b_k\}$ that maximize the difference in predictions between the two theories. We start by defining an optimization function over the space of binnings. We merit partitions that enhance the expected evidence in favor of T_N, $\mathcal{E}(T_N)$,

[3] Recall T is either the new theory T_N or the Standard Model T_S.

if T_N is correct, plus the expected evidence in favor of T_S, $\mathcal{E}(T_S)$, if T_S is correct. Given a particular set of bins, $\{b_k\}_{k=1}^{v}$, the proposed optimization function is defined as follows:

$$\mathcal{O}(\{b_k\}) = \mathcal{E}(T_N, \{b_k\}) + \mathcal{E}(T_S, \{b_k\}) \tag{8}$$

The evidence for each theory is as follows:

$$\mathcal{E}(T_N, \{b_k\}) = \sum_{\lambda_1} \sum_{\lambda_2} \cdots \sum_{\lambda_v} \left(\prod_k \mathrm{P}(\lambda_k | T_N) \right) \times \log_{10} \left(\frac{\prod_k \mathrm{P}(\lambda_k | T_N)}{\prod_k \mathrm{P}(\lambda_k | T_S)} \right) \tag{9}$$

and similarly for $\mathcal{E}(T_S, \{b_k\})$. Each summation on the left varies over the range $[0, \infty]$. The evidence for each theory has a straightforward interpretation. Recall that $\prod_k \mathrm{P}(\lambda_k | T) = \mathrm{P}(D_i | T)$ and therefore each evidence $\mathcal{E}$ is the relative entropy of the data likelihoods (if $\log_{10}$ is replaced with $\log_2$), averaged over all possible outcomes on the number of real events observed on each bin. The two components in equation 8 are necessary because relative entropy is not symmetric. The representation for $\mathcal{O}$ can be simplified as follows:

$$\mathcal{O}(\{b_k\}) = \sum_k \sum_{\lambda_k} (\mathrm{P}(\lambda_k | T_N) - \mathrm{P}(\lambda_k | T_S)) \times (\log_{10} \mathrm{P}(\lambda_k | T_N) - \log_{10} \mathrm{P}(\lambda_k | T_S)) \,, \tag{10}$$

In practice one cannot evaluate $\mathcal{O}$ by trying all possible combinations in the number of real events observed on each bin. Instead we estimate the expected number of events in bin k if theory T is true, $\mu_{k|T}$, and consider $\pm s$ standard deviations (s is user-defined) around that expectation, which can be quickly evaluated with arbitrary accuracy by explicitly computing the sum for those bins with expectation $\mu_{k|T} \leq 25$ and using a gaussian approximation for those bins with expectation $\mu_{k|T} > 25$.

Although in principle maximizing $\mathcal{O}$ requires optimizing the positions of all bin edges simultaneously, in practice it is convenient to choose the bin edges sequentially. Starting with a single bin encompassing all points, this bin is split into two bins at a location chosen to maximize $\mathcal{O}$. At the next iteration, a new split is made that improves $\mathcal{O}$. The algorithm continues iteratively until further division results in negligible or negative change in $\mathcal{O}$. Figure 2 (Algo. 1) illustrates the mechanism behind the binning technique. The complexity of the algorithm is linear in the size of the input space (i.e., in the size of the two datasets $\tilde{D}_{ni}$ and $\tilde{D}_{si}$).

4.2 Example with Gaussians of Varying Width

To illustrate the mechanism behind the bin-allocation mechanism, assume a scenario with two Gaussian distributions of different widths over a variable x. Figure 3(left) shows the true (but unknown) distributions $f_1(x)$ and $f_2(x)$, where $f_i(x) = \frac{n}{\sqrt{2\pi}\sigma_i} e^{(-(x-\mu)^2/2\sigma_i{}^2)}$ with $i = \{1, 2\}$ and parameter values $n = 100$, $\mu = 25$, $\sigma_1 = 5$, and $\sigma_2 = 8$. The units on the vertical axis are the number of events expected in the data per unit x. We used one thousand points randomly drawn from $f_1(x)$ and from $f_2(x)$. These points are shown in the histogram in Fig. 3(right), in bins of unit width in x. The algorithm proceeds to find edges sequentially before halting, achieving a final figure of merit. The resulting bins are concentrated in the regions $x \approx 20$ and $x \approx 30$, where $f_1(x)$ and $f_2(x)$ cross.

```
Algorithm 1: Adaptive-Bin-Allocation
Input: D, $\tilde{D}_{ni}$, $\tilde{D}_{si}$
Output: Set of bins $\{b_k\}$
ALLOCATE-BINS(D,$\tilde{D}_{ni}$,$\tilde{D}_{si}$)
    Evaluate D at each discrete Monte Carlo event in $\tilde{D}_{ni}$ and $\tilde{D}_{si}$.
    Estimate probability densities $f(\mu_{k|T})$ for $T = T_N$ and $T = T_S$.
    Initialize set of bins $\{b_0\}$, where $b_0$ covers the entire domain of D.
    repeat
        Search for a cut point $c$ over D that maximizes function $\mathcal{O}$.
        Replace the bin $b_k$ where $c$ falls with the two corresponding new bins.
    until The value $o^*$ maximizing $\mathcal{O}(\cdot)$ is such that $o^* < \epsilon$
    end
    return $\{b_k\}$
```

Fig. 2. Steps to generate a set of bins that maximize the distance between the events predicted by theory T_N and theory T_S.

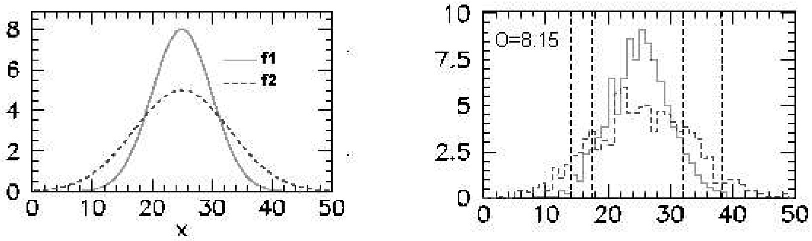

Fig. 3. (Left) Two Gaussian distributions, f_1 and f_2, with same mean but different variance. (Right) The bin-allocation mechanism identifies those regions where f_1 and f_2 cross.

5 Experiments

We describe two examples conducted at the Tevatron ring at the Fermi National Accelerator Laboratory in Chicago, Illinois. The accelerator collides protons and anti-protons at center of mass energies of 1960 GeV (i.e., giga electron volts). A typical real-collision dataset of this collider is made of about 100 thousand events.

We divide each of the Monte Carlo data sets $\tilde{D}_n$, and $\tilde{D}_s$ into three equal-size subsets. The first subset is used to compute the probability densities $\mathrm{P}_i(\mathbf{e}|T_N), \mathrm{P}_i(\mathbf{e}|T_S)$ (Section 3.2); the second subset is used to run the adaptive bin-allocation mechanism (Section 4); the last subset is used to estimate the figure of merit $\mathcal{L}(T_N) = \log_{10} \frac{\mathrm{P}(\mathcal{D}|T_N)}{\mathrm{P}(\mathcal{D}|T_S)}$ (Section 3). Each experiment produces several hundreds of final states. The running time for each experiment was approximately one hour on a Linux machine with a Pentium 3 processor and 1 GB of memory.

Searching for Leptoquark Pair Production. The first experiment is motivated by a search for leptoquark pair production as a function of assumed leptoquark mass. We show how a theory that advocates leptoquarks with small masses –that if true would result in an abundance of these particles compared to their heavier counterparts– is actually disfavored by real data. Figure 4 (left) shows the log likelihood ratio $\mathcal{L}(T_N)$ (equation 1) for different leptoquark masses. Units on the horizontal axis are GeV. The new proposed theory is disfavored by the data for small mass values, but becomes identical to the Standard Model

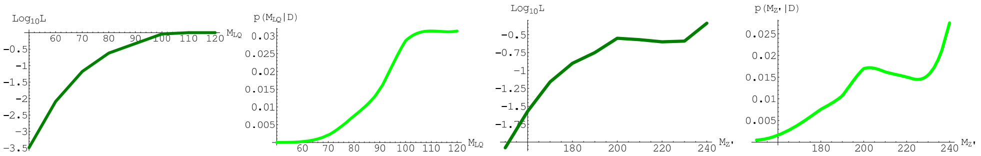

Fig. 4. (left) The log likelihood ratio $\mathcal{L}(T_N)$ (equation 1) for different leptoquark masses. (second left) The posterior distribution $p(M_{LQ}|\mathcal{D})$ obtained from a flat prior and the likelihood on the left. (third left) The log likelihood ratio for different Z' masses. (right) The posterior probability $p(m_{Z'}|\mathcal{D})$ flattens out beyond $m_{Z'} \approx 250$ GeV. Units on the horizontal axis are GeV.

for large mass values. Figure 4 (second left) shows the posterior distribution $p(M_{LQ}|\mathcal{D})$ obtained from a flat prior and the likelihood on the left.

Searching for a Heavy Z' Particle The second experiment is similar in spirit to the previous one. Figure 4(third from left) shows a search for a heavy Z' as a function of assumed Z' mass. Z's with small masses, which would be more copiously produced in the Tevatron than their heavier counterparts, are disfavored by the data. The posterior probability $p(m_{Z'}|\mathcal{D})$ flattens out beyond $m_{Z'} \approx 250$ GeV (Figure 4, right), indicating that the data is insufficiently sensitive to provide evidence for or against Z's at this mass.

6 Conclusions and Future Work

This paper describes an approach to quantify the degree of evidence in favor of a new proposed theory compared to a standard baseline theory. The mechanism adaptively allocates histogram bins that emphasize regions in the variable space where there is a clear difference in the predictions made by the two theories. The proposed mechanism carries two important benefits: 1) it simplifies substantially the current time needed to assess the value of new theories, and 2) it can be used to assess a family of theories by varying a particular parameter of interest (e.g., particle mass).

We expect the procedure outlined here to have widespread application. The calculation of likelihood ratios is common practice in the physical and social sciences; the main algorithm can be easily adapted to problems stemming from other scientific fields. One barrier lies in generating Monte Carlo data to model a theory distribution. Particle physicists have invested huge amounts of effort in producing a detector simulator designed to imitate the behavior of real particle colliders.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grants no. IIS-431130 and IIS-448542.

References

1. Duda R. O., Hart P. E., Stork D. G.: Pattern Classification. John Wiley Ed. 2nd Edition (2001).
2. Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning. Springer-Verlag Ed. (2001).
3. Knuteson, Bruce: Systematic Analysis of HEP collider data. Conference on Statistical Problems in Particle Physics, Astrophysics, and Cosmology. Stanford CA.
4. Kocabas S., Langley P.: An Integrated Framework for Extended Discovery in Particle Physics. Proceedings of the 4th International Conference on Discovery Science, pp. 182-195. Springer Verlag. (2001).
5. Scott D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Series in Probability and Statistics Ed. (1992).