Automatic Signal Enhancement in Particle Physics Using Multivariate Classification and Physical Constraints

R. Vilalta* G. Mutchler† S. Taylor‡ B. Knuteson§

Abstract

We present a methodology to automate the process of signal enhancement in particle physics by relying on multivariate classification techniques. The traditional subjective approach based on visualizing potential peaks and manually isolating them is supplanted by a fully automated process that predicts signal events and exploits physical constraints. Our experiments compare the performance of several multivariate classification techniques (e.g., random forests, Bayesian classification, support vector machines) for signal identification using K* mass as a test case. We show how information about physical constraints obtained from kinematic fitting procedures can be used to enrich the original feature representation. Finally, we suggest a form of hypothesis testing to verify the presence of a particle of interest on real data. Overall, our goal is to provide the particle physics community with computational tools that obviate manual and subjective interpretation of data patterns.

1 Introduction

Experiments in particle physics face the problem of identifying elementary particles produced at frontier energy colliders. Typical colliders have millions of channels of electronics, producing terabytes of data per second. These data are analyzed in real time and reduced to a few terabytes per day that is stored for later analysis. Of the billion particle collisions occurring each second, only a few are of interest. Finding these interesting — but possibly unanticipated — collisions in such a massive data stream represents a challenging test of forefront technology and computational power.

An energetic, strongly-interacting particle entering a calorimeter causes a chain reaction of nuclear breakup and particle production, resulting in a shower of particles passing through the detector. These signals must be reduced to the speed and directions of between roughly 2 and 8 objects emerging from the initial collision for a convenient comparison with theoretical predictions. Substantial processing of the recorded signals from each collision reduces the amount of information to between roughly 10 and 60 features that can be used to determine which of these different objects have produced any particular cluster of energy in the detector. These features can be usefully thought of as forming a *feature space*. The dimensionality of this space and the definitions of each axis in the space depend on the signals recorded in each energy cluster. Examples of variables include calorimeter cluster parameters, number of tracks associated with the event, invariant mass, energy, momentum, spatial coordinates, etc.

1.1 Current Techniques for Particle Identification. Current criteria for identifying each type of elementary particle are crafted by hand at each experiment. These criteria typically take the form of a set of rigid cuts over the feature space. Often each individual cut requires that a particular feature have a value greater than (or less than) some constant. These cuts are determined by simulating how each object looks in the detector, folding in expert knowledge of the underlying physical processes leading to signals in the detector, and trading off errors of Type I (i.e., failing to identify true particles –false negatives) with errors of Type II (i.e., pointing to non-existing particles –false positives). Figure 1 illustrates a particular scenario in the identification of taus ($\tau^{\pm}$). The diagram looks at a specific region of a 2-dimensional feature space populated by these particles. Decision boundaries were manually defined in the form of two horizontal lines and one exponentially decreasing line. No systematic procedure attending to the probability density of this distribution, or the discriminative power against background events, played a role in the definition of these boundaries.

This approach has certain definite advantages. For instance, the application of rigid cuts allows physicists to impose their knowledge of the detector on the identification criteria through an accumulation of additional requirements; requirements are commonly geometrical

*Dept. of Computer Science, University of Houston. 4800 Calhoun Rd. Houston TX 77204-3010. Email: vilalta@cs.uh.edu

†Bonner Nuclear Lab, Rice University. 6100 Main Street, Houston, TX 77005. Email: mutchler@rice.edu

‡Dept. of Physics and Astronomy, Ohio University, 251 Clippinger Labs, Athens, OH 45701. Email: staylor@jlab.org

§Lab for Nuclear Science, MIT. 77 Massachusetts Ave. Cambridge, MA 02139-4307. Email: knuteson@mit.edu

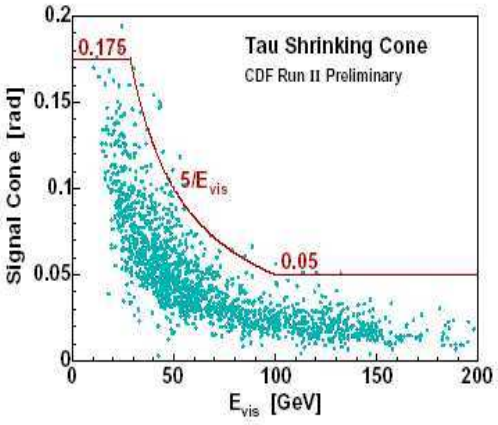

Figure 1: A particular scenario in the identification of taus ($\tau^{\pm}$). Decision boundaries were manually defined in the form of two horizontal lines and one exponentially decreasing line. No systematic procedure attending to the probability density of this distribution, or the discriminative power against background events, played a role in the definition of these boundaries (reproduced from CDF internal note #7399 by Safonov).

in nature. In addition, rigid cuts are conveniently understood visually by considering one-dimensional projections of the full-dimensional space.

But the approach also has built-in limitations:

- No well-defined prescription exists for simultaneously combining the few dozen potentially useful criteria to identify particles, leading to much argument and discussion in collaborations numbering several hundred physicists.
- Cut-based object selection criteria are binary, not easily continuously varied to optimize trade-offs between errors of Type I and errors of Type II. Despite tens of thousands of man-hours spent constructing and improving object identification at each major collider experiment, the resulting criteria are in some cases surprisingly suboptimal (a single definite figure of merit with which to quantify "optimal" does not even exist).

1.2 Objective of this Study. The purpose of this analysis is to gain insight on how to increase the degree of automation in the process of particle identification by exploiting multivariate techniques and physical constraints (and thus avoiding manual and subjective cuts over low-dimensional feature spaces). Traditional techniques that exploit physical constraints use "kinematic fitting" to improve measured quantities and to provide a means to cut background. We propose an additional step where a multivariate classification technique is invoked on Monte Carlo data to generate a predictive model. The model is used to separate signal events from background events. Applying the model to real data results in a (predicted) signal distribution where evidence for the existence of a particle of interest is enhanced. This paper extends previous work on the same topic [1].

2 The Physical Experiment

We begin by describing the physical experiment. A broad band energetic photon beam (γ) hits a liquid hydrogen target, the proton (p). The photon interacts and produces a number of charged and uncharged particles. We look for the following reactions:

$$\gamma \mathrm{p} \rightarrow \Lambda K^{*+} \tag{2.1}$$
$$\gamma \mathrm{p} \rightarrow \Lambda K^{+} \pi^{0} \tag{2.2}$$
$$\gamma \mathrm{p} \rightarrow K^{+} \mathrm{p} \pi^{-} \pi^{0} \tag{2.3}$$

This is all complicated by various background reaction which all have the same measured particles. Background reactions include the following:

$$\gamma p \rightarrow K^{+} \Lambda \tag{2.4}$$
$$\gamma p \rightarrow K^{+} \Lambda^{*} \tag{2.5}$$
$$\gamma p \rightarrow K^{+} \Sigma^{0} \tag{2.6}$$
$$\gamma p \rightarrow K^{+} \Sigma^{+} \pi^{-} \tag{2.7}$$
$$\gamma p \rightarrow K^{+} \Sigma^{-} \pi^{+} \tag{2.8}$$

Our data set contains information about the incident photon (γ), and three charged particles, K^{+}, p, and π^{-}. While the charged particles are detected, the uncharged ones are not seen, and must be inferred from the missing mass (e.g., π^{0}).

For each detected charged particle we measure the momentum p and the polar angle θ and azimuthal angle ϕ. From these quantities we can construct the three vector, $\mathbf{p} = \mathbf{i}p_x + \mathbf{j}p_y + \mathbf{k}p_z$ where $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the unit vectors. We also measure the Time-of-Flight (TOF). From the TOF and momentum we can calculate the mass m of the particle. Finally, for each particle, we are able to construct a 4-vector, $(E, \mathbf{p})$, where $E = \sqrt{p^2 + m^2}$.

In this particular paper we focus on identifying the presence of K^{*+} (equation 2.1) after the photon-proton interaction (γp). This is in practice not of real interest, but stands as a convenient test case to assess the value behind multivariate classification techniques. Invoking these techniques is justified by the inherent difficulty in separating signal events from background events

(many background reactions produce similar measured particles).

2.1 Using Kinematic Fitting and Physical Constraints. Common practice in particle physics is to take advantage of physical constraints such as energy and momentum conservation to improve measured quantities and to provide a means to cut background. We follow that approach through a technique known as kinematic fitting [2]. For this technique we have chosen to use the Lagrange multiplier method. First, the unknown variables are divided into a set of measured variables ($\vec{\eta}$) and a set of unmeasured variables ($\vec{\xi}$) such as the missing momentum or the 4-vector for a decay particle. For each constraint equation a new variable λ_i is introduced. These variables are the Lagrange multipliers. To find the best fit we minimize

$$\chi^2(\vec{\eta},\vec{\xi},\vec{\lambda}) = (\vec{\eta}_0-\vec{\eta})^T V^{-1}(\vec{\eta}_0-\vec{\eta}) + 2\vec{\lambda}^T\vec{f} \tag{2.9}$$

by differentiating χ^2 with respect to all the variables, linearizing the constraint equations and iterating. Here $\vec{\eta}_0$ is a vector containing the initial guesses for the measured quantities, V is the covariance matrix comprising the estimated errors on the measured quantities, and $\vec{f}$ represents the constraints such as energy and momentum conservation.

The solution for iteration $\nu+1$ depends on the result of the previous iteration (ν):

$$\vec{\xi}^{\nu+1} = \vec{\xi}^{\nu} - (F_\xi^T S^{-1} F_\xi)^{-1} F_\xi^T S^{-1}\vec{r}, \tag{2.10}$$

$$\vec{\lambda}^{\nu+1} = S^{-1}[\vec{r} + F_\xi(\vec{\xi}^{\nu+1} - \vec{\xi}^{\nu})], \tag{2.11}$$

$$\vec{\eta}^{\nu+1} = \vec{\eta}_o - V F_\eta^T \vec{\lambda}^{\nu+1}, \tag{2.12}$$

where we define

$$(F_\eta)_{ij} \equiv \frac{\partial f_i}{\partial \eta_j}, \tag{2.13}$$

$$(F_\xi)_{ij} \equiv \frac{\partial f_i}{\partial \xi_j}, \tag{2.14}$$

$$\vec{r} \equiv \vec{f} + (F_\eta)^\nu(\vec{\eta}_0 - \vec{\eta}^{\nu}), \tag{2.15}$$

$$S \equiv (F_\eta)^\nu V (F_\eta^T)^\nu, \tag{2.16}$$

all evaluated at the νth iteration. We iterate until the difference in magnitude between the current χ^2 and the previous value is less than or equal to $\Delta\chi^2_{test}$ (i.e., until the difference is less than 0.001) or the loop has reached its maximum number of iterations (i.e., until the number of iterations is 20).

2.2 Generating Confidence Levels. For our purposes, we are interested in using kinematic fitting to obtain a confidence level (goodness of fit to the data). As an example, let's look into the fitting procedure as applied to the proton (p) and pi-minus (π^-) tracks with the Λ hypothesis. Explicitly, the constraint equations are as follows:

$$\vec{f} = \begin{bmatrix} E_p + E_\pi - E_\Lambda \\ \vec{p}_p + \vec{p}_\pi - \vec{p}_\Lambda \\ (y-y_\pi)p_\pi^z - (z-z_\pi)p_\pi^y \\ (x-x_\pi)p_\pi^z - (z-z_\pi)p_\pi^x \\ (y-y_p)p_p^z - (z-z_p)p_p^y \\ (x-x_p)p_p^z - (z-z_p)p_p^x \end{bmatrix} = \vec{0}. \tag{2.17}$$

The χ^2 distribution for this fit is the result of a fit to the histogram using the functional form of a χ^2 distribution with two degrees of freedom plus a flat background term. Explicitly,

$$f(\chi^2) = \frac{P_1}{2} e^{-P_2\chi^2/2} + P_3. \tag{2.18}$$

where P_1 is simply used for normalization. P_2 is a measure of how close the distribution in the histogram is to an ideal χ^2 distribution (in that case $P_2 = 1$). The equation above (2.18) serves to characterize the behavior of the χ^2 distribution. With two degrees of freedom we expect this distribution to be exponential. However, under the presence of background events from other channels, the distribution may differ; particular, we expect a wide range of values for χ^2 under background events. This is compensated with P_3.

The Confidence Level (CL) is the primary measure of the goodness of fit to the data and is given by the equation

$$CL = \int_{\chi^2}^{\infty} f(z;n)dz \tag{2.19}$$

where f(z:n) is the χ^2 probability density function with n degrees of freedom (where we have assumed normally distributed errors).

3 Using Multivariate Classification Techniques

In addition to the traditional approach of kinematic fitting, we suggest using multivariate classification techniques for signal identification and enhancement. Our approach consists of using the confidence levels (goodness of fit to the data described above) as new features into a classification problem. The resulting model implicitly uses kinematic-fitting results to further enhance the signal of interest (e.g., to enhance K^{*+}).

3.1 The Classification Problem. We begin by giving a brief overview of the classification problem [3, 4]. A classifier receives as input a set of training examples $T = \{(\mathbf{x}, y)\}$, where $\mathbf{x} = (a_1, a_2, \cdots, a_n)$ is a vector or point in the input space ($x \in \mathcal{X}$), and y is a point in the

Table 1: Columns 2-3: Mean accuracy performance (Acc.) with different misclassification costs. Numbers enclosed in parentheses represent standard deviations. Columns 4-5: Mean false positive rates (FPR) with different misclassification costs.

Analysis Technique	Acc. Equal Costs	Acc. Unequal Costs	FPR Equal Costs	FPR Unequal Costs
Naive Bayes	85.59 (0.86)	86.79* (0.78)	20.1	6.8
Support Vector Machines	87.69 (0.70)	88.29 (0.51)	18.7	1.6
Multilayer Perceptron	88.57 (0.85)	90.58 (0.73)	14.3	3.0
ADTree	88.90 (1.14)	90.81* (0.96)	11.5	3.7
Decision Tree	89.23 (0.93)	91.97* (0.87)	12.7	4.7
Random Forest	90.02 (1.12)	92.34* (0.95)	11.6	4.3

output space ($y \in \mathcal{Y}$). We assume T consists of independently and identically distributed (i.i.d.) examples obtained according to a fixed but unknown joint probability distribution. The outcome of the classifier is a function h (or hypothesis) mapping the input space to the output space, $h : \mathcal{X} \rightarrow \mathcal{Y}$. Function h can then be used to predict the class of previously unseen attribute vectors.

3.2 Data for Analysis. In our study, the output variable for each event indicates if the photon-proton interaction resulted in the production of K^{*+} (positive event) or not (negative event). Each feature vector $\mathbf{x}$ is made of 45 features. The first 4 features are confidence level numbers derived from the kinematic fits (Section 2.2). The next feature corresponds to the total energy. The last 40 features characterize 8 particles (3 of them detected and 5 inferred). Each particle is represented by energy E, momentum p, polar angle θ, azimuthal angle ϕ, and mass squared m^2.

Our data set is derived using the CEBAF large angle spectrometer (CLAS). We gathered 1000 Monte Carlo signal events and 6000 Monte Carlo background events. The real data comprised about 13,500 events.

3.3 Using Monte Carlo Data and Variable Misclassification Costs. Our first set of experiments were limited to Monte Carlo data for which the value of the output variable of each event is known. Our study compared the performance of several classification algorithms in terms of predictive accuracy. We employed several algorithms including decision trees, support-vector machines, random forests, etc. We used the machine-learning tool WEKA [5] with default parameters for all learning algorithms.

First we reduced the original size of the input space through a feature selection process, using information gain as the evaluation metric [4]. The number of features was reduced from 45 to 5 (the final set included information on the confidence levels). For each algorithm we varied the amount of misclassification costs. Table 1 shows our results. The first column describes the multivariate classification techniques used for our experiments. The second column shows accuracy estimations with equal misclassifications costs; the third column shows accuracy estimations where the cost of a false positive is 3 times more expensive than the cost of a false negative. Each result is the average of 5 trials of 10-fold cross validation each [4]. An asterisk at the top right of a number implies the difference is significant at the $p = 0.01$ level (assuming a two-tailed t-student distribution). Overall there is a significant increase in performance by adding a penalty when mislabelling background events as target events. In addition, Table 1 shows how for this particular domain, varying misclassification costs can yield a significant reduction in the false positive rate (FPR %, columns 4-5).

Our results denote a preference for the strategy behind "random forests". We have observed similar results in other experiments [6]. Random forests have the ability to reduce the variance and bias components of error by voting over multiple decision trees using on each tree a random selection of features [7]. They exhibit robust behavior against problems with multiple systematic errors as is common to problems in particle physics.

3.4 Signal Enhancement on Real Data. Our next set of experiments used real data for which the value of the output variable of an event is unknown. In this case the problem is not to maximize accuracy performance (i.e., minimize a risk functional such as zero-one loss) but instead to provide enough evidence to believe that the signal event occurred multiple times during the photon-proton interaction. The goal is to find a technique able to enhance the signal distribution

over the background distribution.

Our approach to deal with the signal enhancement problem is as follows. Applying a multivariate technique M on Monte Carlo data yields a predictive model h_M. One can then apply h_M on the real data to generate a histogram for the predicted signal distribution. If model h_M exhibits good performance, we expect the histogram generated through h_M to provide evidence for the occurrence of the desired signal.

To illustrate our approach Figure 2 (upper) shows a histogram generated with all real data; the x-axis corresponds to the squared mass (m^2) of the signal particle (K^{*+}). Figure 2 (middle) shows a histogram generated by taking only those events predicted as signal on the real data by a classification model. Kinematic fitting variables were part of the feature vectors. We employed random forests as the classification technique; the derived information helps isolate and enhance the signal distribution. Figure 2 (lower) shows the corresponding histogram using random forests with cost sensitive classification and kinematic fitting variables. The resulting histogram shows an even larger enhancement over the signal distribution.

3.5 Measuring the Quality of Signal Enhancement. Lastly, we face the problem of assessing the quality of the signal enhancement process. Traditionally, a form of hypothesis-testing is used to determine if the bump on Figure 2 (upper) can be interpreted as coming from a distribution other than the background distribution. We claim this approach has several limitations as explained next.

The classical methodology works as follows. The real data is fit by a Breit-Weigner function together with a background function [8]. One possibility for the combined signal-background function is as follows:

$$f(x) = \mathrm{BW}(x) + \mathrm{BG}(x) \tag{3.20}$$

where BW(x) is the Breit-Weigner function (similar to a Gaussian distribution):

$$\mathrm{BW}(x) = \frac{(\Gamma/2)}{(x - X_0)^2 + (\Gamma/2)^2} \tag{3.21}$$

where X_0 and Γ are parameters specifying the mean and width of the distribution. The background function is defined as follows:

$$\mathrm{BG}(x) = a(1 + b(x - \mu) + c(x - \mu)^2) \tag{3.22}$$

where a, b, and c are parameters describing the amount of background, shape and curvature of the distribution. The fitting is done over a user-defined range $[X_i, X_f]$ with a mean of μ. In our case consider x as the squared mass; equation 3.20 can then be used to model the histogram for real data (Figure 2 upper). After estimating all parameters (e.g., using maximum likelihood estimation) a a figure of merit responds to the question: what is the probability that this background (as represented by the polynomial) fluctuated and gave rise to this Breit-Wigner distribution?.

A clear limitation of the traditional approach is the subjectivity with which the data is used to decide the focus of the experiment. The range $[X_i, X_f]$ over which the analysis is done is decided after looking at the data. In this case the sampling error goes unaccounted for. In addition, an optimization technique is needed to estimate the value of many parameters simultaneously.

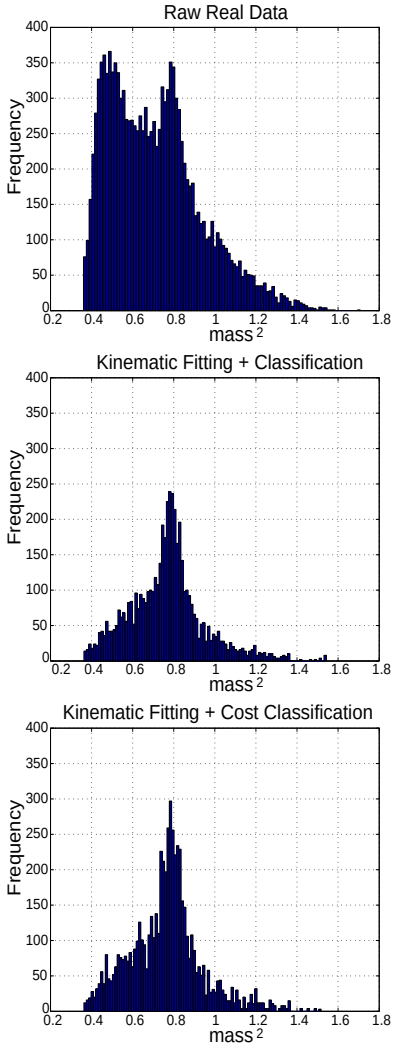

Figure 2: Histograms using (upper) real data (middle) predicted signals on real data by random forests, and (lower) predicted signals on real data by random forests using cost-sensitive information. The x-axis corresponds to K^{*+} squared mass (units are in $\frac{\mathrm{GeV}^2}{c^4}$).

Rather than following the traditional approach, we propose a two step process. The first step is simply to visually compare the difference between Figure 2 (middle) and Figure 2 (lower) with respect to the hypothetical form of the signal distribution as dictated by Monte Carlo data (assuming the Monte Carlo adequately models the background and signal). This can serve as a fast way to verify that our empirical distributions look close to the hypothetical true signal distribution (see Figure 3 lower). In our case both signal distributions on real data look very similar to the hypothetical true signal.

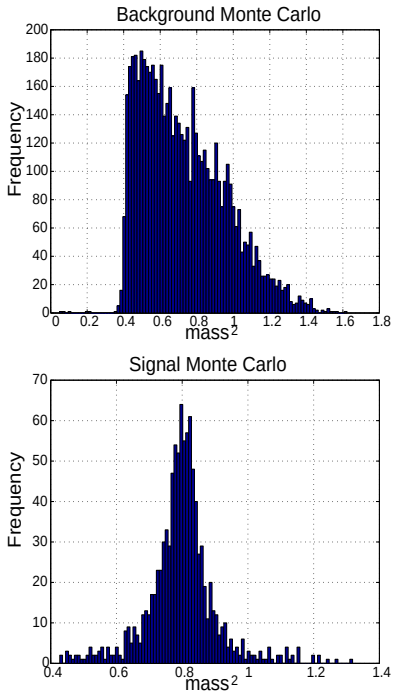

Figure 3: Histograms using Monte Carlo data for (upper) background distribution and (lower) Signal Distribution. These are hypothetical distributions assuming the Monte Carlo adequately models the background and signal. The x-axis corresponds to K^{*+} squared mass (units are in $\frac{\text{GeV}^2}{c^4}$).

The second step is to compare the distribution of error on both the Monte Carlo data and the real data using our classification model. The rationale is as follows. If the real data is made of only background events and does not contain the particle under study, we would expect the proportion of events predicted as signal in the real data, ψ, to be similar to the false positive rate, δ, found when the model was tested on Monte Carlo data. This rate would simply stand as the inherent inaccuracy exhibited by the model. On the other hand, the presence of the particle under study in real data would generate a proportion of events predicted as signal ψ higher than that of δ (i.e., we would expect $\psi > \delta$). One could approximate these rates by repeated experiments to come up with estimates of their means (μ_δ,μ_ψ) and standard deviations (σ_δ, σ_ψ) for both error distributions. A form of hypothesis testing can then be applied with the null hypothesis asserting $H_0 : \delta = \psi$ and the alternative hypothesis asserting $H_1 : \delta < \psi$.

We applied the step described above to our data. For the Monte Carlo data we used a form of random subsampling by randomly dividing the data into a training set (using 90% of the data) and a testing set (using 10% of the data); we repeated this 30 times to yield error estimates for μ_δ = and σ_ψ. In the case of real data we tested a classification model (random forests) on a random sample corresponding to 10% of the data and repeated this 30 times to obtain estimates for μ_δ and σ_ψ. Assuming a Gaussian distribution for both error models we obtained the following results. Using random forests with no cost sensitive classification the null hypothesis is rejected with a confidence value of $p = 0.09$ (i.e., the probability of rejecting the null hypothesis when it is in fact true is 9%); using random forests with cost sensitive classification the null hypothesis is rejected with a confidence value of $p = 0.19$ (i.e., the probability of rejecting the null hypothesis when it is in fact true is 19%). These are reasonable values that suggest the presence of K^{*+} on the data. Our proposed methodology thus provides a way to determine the presence of a particle of interest (with a confidence value) without the problem of manually identifying regions of interest over the mass histogram.

4 Conclusions and Future Work

Our study exploits information derived from physical constraints in the forms of confidence levels (using kinematic fitting) to enrich the set of original features for classification. We suggest generating a predictive model over Monte Carlo data to produce a distribution over real data where a signal of interest is enhanced. Our model integrates information about physical constraints and exploits classification techniques to automate the process of signal identification and enhancement.

One unexplored area is to determine the degree to which multivariate classification techniques contribute to signal enhancement without any information derived from kinematic fitting. It is important to understand how current classification techniques can exploit information derived from physical constraints.

Our approach obviates forming decisions "a posteriori" about the interestingness of a bump signal protrud-

ing above a background distribution; this is important to avoid the risk of finding patterns stemming from random sampling. Our study simply shows how a classification model trained "a priori" on Monte Carlo data can be used to identify possible signal events. Our methodology performs a form of hypothesis testing comparing the distribution of error rates. Future work will address how to compare both distributions attending to additional properties (e.g., mean, variance, skewness, kurtosis, etc.).

Acknowledgments

Thanks to Abraham Bagherjeiran for helping to compute the relative distance function and to Puneet Sarda for carrying the experimental comparison of learning algorithms. This material is based upon work supported by the National Science Foundation under Grants no. IIS-431130 and IIS-448542, and by the Department of Energy under Grant no. DE-FG03-93ER40772.

References

[1] R. Vilalta, P. Sarda, G. Mutchler, P. Padley, S. Taylor (2005). *Signal Enhancement Using Classification Techniques and Physical Constraint*, Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT05).

[2] A. G. Frodesen (1979). *Probability and Statistics in Particle Physics*, Oxford University Press.

[3] R. O. Duda, P. E. Hart, and D. G. Stork (2001). *Pattern Classification*, John Wiley Ed. 2nd Edition.

[4] T. Hastie, R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer-Verlag.

[5] I. H. Witten and E. Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press, London U.K.

[6] P. Bargassa, S. Herrin, S-J Lee, P. Padley, R. Vilalta (2005). *Application of Machine Learning Tools to Particle Physics*, Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT05).

[7] L. Breiman (2001). *Random Forests*, Machine Learning 45(1) pp. 5-32. Springer Science-Business Media.

[8] L. Lyons (1986). *Statistics for Nuclear and Particle Physics*, Cambridge University Press.