Statistics for the LHC: Progress, Challenges, and Future

Kyle S. Cranmer
New York University

Abstract

The Large Hadron Collider offers tremendous potential for the discovery of new physics and poses many challenges to the statistical techniques used within High Energy Physics. I will review some of the significant progress that has been made since the PhyStat 2005 conference in Oxford and highlight some of the most important outstanding issues. I will also present some ideas for future developments within the field and advocate a progressive form of publication.

1 Introduction

There are several direct and indirect indications that some type of new physics will show up at the TeV scale – the energy scale being explored by the Large Hadron Collider (LHC) and the multi-purpose detectors ATLAS and CMS. There are a plethora of theoretical models that have been proposed for this new physics; some with few parameters that make specific predictions, some with many parameters and diverse phenomenology[1], and some that are quite vague. For the models that make sharp predictions, I will summarize the substantial progress that has been made recently regarding the statistical procedures used to establish a discovery, and indicate some of the challenges and open issues that remain. In the case of high-dimensional models or models with vague predictions, the statistical challenge is more strategical in nature. I will outline some of the approaches that have been proposed and discuss some new directions that may bear fruit. In addition to the work being done by experimentalists, I will review some of the work being done by the growing community of theorists using sophisticated statistical techniques. I will conclude with some discussion of how the theoretical and experimental communities can improve their communication and speed the iteration cycle needed to interpret signs of new physics.

This paper is largely a continuation of my contribution to PhyStat 2005, and I urge readers to consult those proceedings for a more thorough statement of the problem and introduction to notation [1]. For completeness, it should be said that within High Energy Physics (HEP) we use a theoretical formalism called Quantum Field Theory that allows us to predict the "cross-section" of any particular interaction, which is proportional to the probability that it will occur in a given collision. The number of observed events, n, is Poisson distributed. Distributions of discriminating variables (angles, energies, masses, etc.) of the particles produced in a collision are described as a convolution of fundamental distributions predicted by theory and complicated detector effects that can only be modeled with Monte Carlo techniques. The resulting distributions are generically called "shapes" and are denoted $f(m)$ (where m may have many components). The "Standard Model" is a specific theory that has survived all our tests so far, thus it is our Null, or "background-only", hypothesis. Uncertainties in the detector performance, deficiencies in our theoretical modeling, and finite computational resources lead to uncertainties in our prediction of the background and often force us to resort to an effective description using a parametric model $f(m|\nu)$ instead of Monte Carlo. The parameters ν are nuisance parameters and reflect our uncertainty in the background. Incorporating nuisance parameters into (or eliminating them from) the statistical techniques used to claim discovery is the focus of the next section. The situation is complicated when the signal hypothesis is composite (has additional free parameters). When there are relatively few parameters in the signal model (eg. $\{m_H\}$ or $\{m_A, \tan\beta\}$) we refer to the problem as the "look elsewhere effect". A more severe form of this problem occurs when the model space has many parameters (e.g. $\{m_0, M_{1/2}, A_0, \tan\beta\}$, the 105 parameters of the MSSM, or the even larger set of models in hep-ph), calling for a more radical approach.

[1] Phenomenology in this context refers to the expected signature of new physics.

2 Searches for Specific Signatures

The claim of discovery of new physics is a statement that the data are inconsistent with our current Standard Model to a high degree. Often, the signature of new physics is evidence of an excess in some distribution above the background. Compared to recent experiments, the LHC experiments have a combination of large background uncertainties and an enormous discovery potential. The large background uncertainties are largely due to the fact that the machine collides protons, which are not fundamental objects, and because we will probe new kinematic regimes. There has been a noble effort by the theoretical community to model these effects and improve Monte Carlo tools; however, it is expected that there will still be significant uncertainties in the rate and shape of the various backgrounds to new physics searches [2, 3].

The expected number of events from background processes is typically denoted b, and b is used as a subscript when needed.[2] Hence the model for our null hypothesis has the form of a "marked Poisson" and can be written

$$L(\mathbf{m}|H_0) = \text{Pois}(n|b) \prod_j^n f_b(m_j;\nu), \tag{1}$$

where ν represents the nuisance parameters used to incorporate uncertainty in our background model and the boldface $\mathbf{m}$ is used to indicate we have a measurement of m for each of the n events.[3] Similarly, the signal is often manifest as an excess above the background, with an expected rate and shape, denoted s and $f_s(m)$. Thus when the signal is purely additive, the model for the alternate hypothesis can be written

$$L(\mathbf{m}|H_1) = \text{Pois}(n|s+b) \prod_j^n f_{s+b}(m_j;\nu) = \text{Pois}(n|s+b) \prod_j^n \frac{s f_s(m_j) + b f_b(m_j;\nu)}{s+b}. \tag{2}$$

When the signal is not additivie (eg. in cases like the Z' where interference effects lead to a deficit) the shape for the alternate, $f_{s+b}(m_j)$, is not a simple mixture model. Often the signal model also has free parameters, but that complication is deferred to Section 2.4.

For quite some time, High Energy Physics has been aware of the Neyman-Pearson lemma and heavily utilized the event-wise likelihood ratio $L(m_j|H_1)/L(m_j|H_0)$ for the selection of signal candidates or the experiment-wise likelihood ratio $L(\mathbf{m}|H_1)/L(\mathbf{m}|H_0)$ as a test statistic in hypothesis testing [4]. The main area of development in the last few years has been the treatment of the nuisance parameters ν and uncertainty in the background rate b [1, 5, 6, 7, 8, 9].

In the LEP Higgs searches, background shapes were known quite well, and shape uncertainties were essentially neglected – or, more accurately, were treated as a systematic error in a way that was decoupled from the rest of the statistical formalism. Normalization uncertainties were included into an otherwise frequentist calculation by "smearing" the background rate according to some (posterior) distribution. This technique of smearing is fundamentally Bayesian (via integrating or marginalizing the nuisance parameter b), and is referred to by several names including Prior Predictive, Cousins-Highland, Z_N, S_{cP}, etc [10, 11, 12]. Searches at the Tevatron have had to deal with shape uncertainties, and the `MCLimit` program developed by Tom Junk employs a mixture of integration and maximization to eliminate nuisance parameters. The techniques being used by the Tevatron currently appear to be adequate for relatively low significance statements (eg. 2σ limits), but may not have good coverage properties at high significance (eg. the 5σ customary for discovery).

2.1 Number Counting Experiments

Analyses that do not take advantage of shape information are called number-counting analyses, and rely purely on the Poisson nature of the counts. Because there are no other discriminating variables, the

[2] With the exception of s and b, Roman characters are reserved for observable quantities and Latin characters are used for model parameters

[3] L will be used interchangeably for a probability density function and a likelihood function

role of background uncertainty is of utmost importance. There has been considerable attention paid to a prototype problem in which a subsidiary measurement y is used to constrain the background rate and the main measurement x is used to test the presence of signal [11].

$$L(x, y|s, b) = \text{Pois}(x|s + b)\text{Pois}(y|\tau b). \tag{3}$$

In my contribution to PhyStat2005, I compared the coverage of several methods in High Energy Physics for calculating significance. The surprising result from that study was that the Bayesian smearing technique, one of the most common techniques in HEP, significantly undercovered for $b = 100$ and $\tau = 1$, an important regime for the LHC. This result was generalized by Cousins and Tucker [12].

An encouraging result of the study presented in my PhyStat2005 contribution was that the use of the Profile Likelihood Ratio[4]

$$\lambda(s = 0) = \frac{L(x, y|s = 0, \hat{\hat{b}})}{L(x, y|\hat{s}, \hat{b})} \tag{4}$$

together with the assumption that $-2 \log \lambda$ is distributed as χ^2_1 (under the null) had good coverage out to 5σ.[5] It is somewhat surprising that the asymptotic result worked so well even with a single observation (x, y) and modest values of the background rate (the Poisson parameter b). This result has spurred significant interest in use of the Profile Likelihood Ratio for LHC searches since it is capable of dealing with many nuisance parameters, has good coverage properties, and can be implemented with one of our field's most throughly debugged tools: `MINUIT/MINOS` [13].

2.2 Coverage Studies With Shapes

In order to explore the coverage properties of the Profile Likelihood Ratio in the presence of shapes and nuisance parameters associated with the shapes, Jan Conrad and I performed a massive Monte Carlo coverage study. We considered a simple extension of the prototype problem:

$$L(x, y, \mathbf{m}|s, b, \nu) = \text{Pois}(x|s + b)\text{Pois}(y|\tau b) \prod_{j=1}^{x} \frac{s f_s(m_j|\nu) + b f_b(m_j|\nu)}{s + b} \tag{5}$$

where

$$f_s(m|\nu) = \frac{1 - e^{-\nu}}{\nu} e^{\nu(m-1)} \qquad \text{and} \qquad f_b(m|\nu) = \frac{1 - e^{-\nu}}{\nu} e^{-\nu m}.$$

We generated $O(10^8)$ pseudo experiments for several values of s, b, and ν. For each we used `MINUIT` to fit $L(x, y, \mathbf{m}|s = 0, \hat{\hat{b}}, \hat{\hat{\nu}})$ and $L(x, y, \mathbf{m}|\hat{s}, \hat{b}, \hat{\nu})$, and used the asymptotic distribution $-2 \log \lambda \sim \chi^2_1$. We tested coverage for background-only scenarios and signal-plus-background scenarios when the shape parameter was assumed to be known and when it was a nuisance parameter. In each of the cases we studied, the coverage from assuming the asymptotic distribution of $-2 \log \lambda$ was very good. There was some indication that for strong discrimination in the shape (large values of $|\nu|$) that there was some undercoverage. See Tab. 1 for the results of that study. While we had planned to compare the power of the profile likelihood ratio with the Bayesian marginalization technique, that study has not been concluded, partially due to the fact that the Bayesian calculation is much more computationally intensive.

This scenario is somewhat artificial because the nuisance parameter ν is shared between the signal and background contributions. Often, the signal shape has its own nuisance parameters, ν', and those nuisance parameters have no effect on the likelihood when $s = 0$. This results in a non-χ^2 distribution for $-2 \log \lambda$ and requires special care. The look-elsewhere effect is an example of this situation, and one must either rely on another asymptotic distribution, use Monte Carlo to estimate the distribution, or recalibrate the p-value obtained.

[4] The use of a single $\hat{}$ denotes the unconditional maximum likelihood estimate, while the double $\hat{}$ denotes the conditional maximum likelihood estimate under the constraint $s = 0$.

[5] If one constrains $s > 0$, then one expects $-2 \log \lambda \sim 1/2\delta(0) + 1/2\chi^2_1$. The factor of 2 in the p-value has a small influence in the significance expressed in σ when one is testing at the 5σ level.

Table 1: Performance of a 5σ test using the profile likelihood ratio in a simple model including shapes with nuisance parameters. Coverage is expressed in σ and quantifies the probability that the true s was included in the 5σ confidence interval. Power is the probability to reject the $s = 0$ hypothesis at the 5σ level.

s	b	τ	ν	coverage [σ]	power [%]
0	20	1	-1	5.1	-
0	40	4	-4	5.1	-
25	100	1	-1	5.1	1.4
50	100	1	-1	5.0	12
50	100	1	-3	4.8	99

2.3 Combining Search Channels

It is common that a new physics signature is manifest in multiple different particle interaction processes. For instance, if the Higgs boson exists, it is expected to decay into different combinations of final state particles with different rates. When the final state particles are different, a different analysis is required: these are commonly referred to as "channels". Obviously, we have more sensitivity to the new physics signature if we combine the different channels. Combining multiple channels by considering the likelihood ratio for the multiple-channel experiment as a test statistic was used by the LEP Higgs searches and is a widely accepted technique.

The LEP Higgs group combined multiple channels by using the multi-channel likelihood ratio

$$\frac{L(\mathbf{m}|H_1)}{L(\mathbf{m}|H_0)} = \prod_{i\in\text{channels}} \left[\frac{\text{Pois}(n_i|s_i+b_i)\ \prod_j^{n_i} \frac{s_i f_{s,i}(m_{j,i})+b_i f_{b,i}(m_{j,i})}{s_i+b_i}}{\text{Pois}(n_i|b_i)\ \prod_j^{n_i} f_{b,i}(m_{j,i})} \right]. \tag{6}$$

It should be noted that here the alternate is a simple model where each of the s_i are known. Moreover, this implies that the relationship of the s_i's is known and is incorporated in the discrimination with the null hypothesis.

As previously mentioned, uncertainty in the background rate, b_i, was included by marginalizing it with respect to some distribution $P(b_i)$, which was taken as a truncated Gaussian and can be considered as a posterior distribution for b_i. The ATLAS experiment used the same formalism to calculate its sensitivity to a low-mass Higgs boson with a pure number counting analysis (eg. no use of f_s or f_b) [14]. Given the undercoverage that was found in this technique of incorporating background uncertainty [1, 12], the combination was repeated using the profile likelihood ratio. In that case, each channel's likelihood function was extended to include an auxiliary, or sideband, measurement y_i that constrains the background via $\text{Pois}(y_i|\tau_i b_i)$. In order to maintain the structure between the different channels (eg. keeping constant the ratios $s_i/s_{i'}$) the s_i were considered to be fixed and an overall signal strength, μ, (related to the production cross-section of the particle) was introduced. Thus, the multi-channel model

$$L(\mathbf{x},\mathbf{y}|\mu,\mathbf{s},\mathbf{b}) = \prod_{i\in\text{channels}} \text{Pois}(x_i|\mu s_i + b_i)\,\text{Pois}(y_i|\tau_i b_i), \tag{7}$$

and the profile likelihood ratio

$$\lambda(\mu=0) = \frac{L(\mathbf{x},\mathbf{y}|\mu=0,\mathbf{s},\hat{\hat{\mathbf{b}}})}{L(\mathbf{x},\mathbf{y}|\hat{\mu},\mathbf{s},\hat{\mathbf{b}})} \tag{8}$$

was used as the test statistic for the combination. See Tab. 2 for a comparison of profile likelihood ratio combination and the marginalization performed with `LEPStats4LHC` [17].

Rolke and López considered the same combination, but used two different approaches [18]. In the technique they called MaxLRT, they considered a discovery to be determined not by the combined likelihood ratio, but solely by the most significant channel. In order to re-calibrate the p-value they

Table 2: Comparison of expected significance of ATLAS Higgs searches with 5 fb^{-1} of data calculated with Prior-Predictive and Profile-Likelihood Ratio. Note, this table is based on previously published ATLAS estimates, but is not itself a result of the ATLAS collaboration. The table is intended to draw attention to the relative difference of the methods rather than the expected significance.

m_H (GeV)	Smearing [σ]	Profile [σ]
110	2.11	1.83
120	3.45	2.43
130	4.76	3.83
140	6.78	5.21
150	8.78	7.45
160	10.43	9.92
170	10.19	9.65
180	8.57	8.02
190	5.77	5.57

made a Bonferroni-type correction. The technique they called FullLRT considered a likelihood ratio as a product of each of the individual channels, but without the notion of an overall signal strength μ. Instead, they took

$$L_{Full}(\mathbf{x},\mathbf{y}|\epsilon,\mathbf{s},\mathbf{b}) = \prod_i \mathrm{Pois}(x_i|\epsilon_i s_i + b_i)\,\mathrm{Pois}(y_i|\tau_i b_i).$$

The p-value for their method is based on the distribution of $-2\log\lambda$ being a linear combination of χ^2 distributions. The main difference in this approach to what was considered in Eq. 8 is that the ratio of unconstrained maximum likelihood estimators $\hat{\epsilon}_i\hat{s}_i$ are not constrained to have the same structure as $\hat{\mu}s_i$. It seems intuitively obvious to me (as a consequence of the Neyman-Pearson lemma) that imposing the additional structure assumed by the alternate hypothesis will translate to additional power, but this has not been confirmed explicitly. Thus, it remains an **open question** if Eq. 8 is more powerful than L_{Full}.

2.4 The Look Elsewhere Effect

So far we have considered the scenario in which the signal model is well specified, and focused on the incorporation of the nuisance parameters ν in the background model. The coverage property that we want our to satisfy is that the rate of Type I error is less than or equal to α *for all* values of the nuisance parameter (eg. $\forall\nu\ \alpha(\nu) < \alpha$). Geometrically, the discovery region corresponds to the *union* of the acceptance regions at every ν, and this union may cause over-coverage for any particular ν.

Now consider the case in which the signal is composite. Let us separately consider signal parameters with physical significance, γ, and those which are more akin to background nuisance parameters, ν_s. If one is interested in γ, then it is not a nuisance parameter and we should expect to represent our results in the $s-\gamma$ (or $\mu-\gamma$) plane. Consider for a moment that γ corresponds to the true mass of the Higgs boson, then our results would be reported in terms of contours in the Higgs cross-section and mass plane. In that case, for every point in the plane, we are asking if the data are consistent with that particular point in the plane. If we restrict ourselves to questions of this form, there is no problem and the relationship between Frequentist confidence intervals and inverted hypothesis tests is clear.

A problem does arrise, however, if one makes a claim of discovery if there is an excess *for any* value of the signal parameter γ (eg. for any mass of the Higgs boson). Clearly, we have a much larger chance to find an excess in a narrow window if we scan the window across a large spectrum. This is often called the "look elsewhere effect", and is typically corrected by scaling the p-value by the "number of places that we looked" or a "trials factor" (often the mass range divided by the mass resolution). This approach of scaling the observed p-value by the trials factor is often called a Bonferroni-type correction.

Formally, one might write "discovery!" $\iff \exists\gamma \ni \forall\nu\ p(\nu,\gamma) < \alpha$. Geometrically, the discovery region now corresponds to *intersection* of the acceptance regions across γ, and this intersection may cause under-coverage for all γ.

In the context of the profile likelihood ratio, one does not explicitly scan γ, but implicitly scans when finding the maximum likelihood estimate $\hat{\gamma}$. The look elsewhere effect is manifest by a non-χ^2 distribution for $-2\log\lambda$. This is known to happen in cases where the alternate model has parameters that do not belong to the null hypothesis (eg. the mass of a new particle sitting on a smooth background). At this conference, Luc Demortier presented various modified asymptotic distributions for $-2\log\lambda$ in these cases [15]. Furthermore, Bill Quayle demonstrated the non-χ^2 distributions via Monte Carlo simulation and proposed an insightful technique to estimate the distribution [16]. It is worth mentioning that the conditions that lead to non-χ^2 distributions for $-2\log\lambda$ seem to only be relevant for discovery, and that all the conditions for a χ^2 distribution are satisfied for measurements of γ or even setting limits on s.

For simple cases (eg. when γ is 1-dimensional), the Bonferroni-type correction is quite straightfoward. In Section 4, I will consider cases in which γ is high-dimensional and the trials factor is either difficult to estimate or leads to a significant loss of power. It is an **open question** whether in simple cases the look elsewhere effect "factorizes" in the sense that a simple re-calibration of the "local" p-value has the same power compared to a method that incorporates the look elsewhere effect in the distribution of the test statistic. A counter example would be equally helpful. Another **open question** is what effect other nuisance parameters in the signal ν_s have on the asymptotic distributions of $-2\log\lambda$.

2.5 Coverage as Calibration & Comparing Multiple Methods

In the last six months, both ATLAS and CMS have created their own statistics committees, and we have already convened joint ATLAS-CMS statistics sessions. One of the outcomes from those discussions was that we plan on using multiple methods for computing the significance of a (hopefully) future observation, and for incorporating systematic errors. As was shown in Ref. [1], the true rate of Type-I error from the different methods may deviate significantly from the nominal value (eg. over- or under-coverage). While one may argue that the accuracy of the coverage at 5σ is irrelevant in absolute terms, it is quite important in relative terms. In particular, we do not want to be in a situation where one experiment requires substantially less data to make a claim of discovery if it is purely due to convention and not because it is actually more powerful. This has furthered the notion that one can think of coverage as a way to "calibrate our statistical apparatus".

Developments such as the RooFit/RooStats framework are being developed to allow us to easily compare different techniques (eg. methods based on the Neyman-Construction, the "profile" construction, profile likelihood ratio, and various Bayesian methods) within the same framework [43].

3 Some Comments on Multivariate Methods

As mentioned in Sec. 2, our field has has been aware of the Neyman-Pearson lemma and heavily utilized the event-wise likelihood ratio $L(m_j|H_1)/L(m_j|H_0)$ for the selection of signal candidates. Here we remind the reader that m_j may be a multi-component discriminating variable and introduce the index k for those d components. To avoid clutter, the event index j will be suppressed. The most basic multivariate analysis, often called "naive Bayes", ignores correlations among the components and builds the event-wise likelihood as a simple (naive) product, viz. $L(m|H_0) = \prod_{k=1}^{d} L(m_k|H_0)$. This technique is very common within HEP, but is rapidly being displaced by other multivariate classifiers like neural networks, decision trees, etc. that can incorporate and leverage non-trivial correlations. It is not surprising that those multivariate classifiers have better performance; however, there are often objections to their "black-box" nature. Furthermore, it is less clear how to incorporate the systematic uncertainties of the Monte Carlo procedures that produced the training data used to train these classifiers. Finally, many of the classifiers have been borrowed from computer science and are optimized with respect to classification

accuracy, a GINI index, or some other heuristic that may not be the most appropriate for the needs of HEP. The next two subsections consider two aspects to multivariate analysis that I hope will complement the other contributions in these proceedings.

3.1 Optimization

Within the context of a search for new physics, one wants to optimize the power of an experiment. In an earlier PhyStat contribution, I introduced the notion of direct and indirect multivariate methods [19, 20]. Essentially, direct methods, such as the genetic programming approach introduced to HEP in Refs. [21, 22], attempt to directly optimize a user-defined performance measure – in this case the power of a 5σ search including background uncertainty. It was shown that many of the common heuristics lead to a function that is at least approximately one-to-one with the likelihood ratio. When neglecting background uncertainty the background hypothesis is no longer composite, both the null and alternate hypotheses are simple, and the Neyman-Pearson lemma holds; thus, in those cases optimization with respect to the heuristic coincides with optimization of power. However, when background uncertainty is taken into account, it is no longer obvious if the heuristic is actually optimizing the power of the search.

A similar point was made by Whiteson and Whiteson when they compared neural networks optimized for classification accuracy to neural networks that were directly optimizing the uncertainty in a top mass measurement [23]. Intuitively, they realized that the top mass measurement is more sensitive to some backgrounds than others, so classification accuracy missed an essential aspect of the problem they were trying to solve. In that case, they found that the uncertainty on the top mass measurement for the classifier that was directly optimizing the mass measurement was $\sim 29\%$ smaller than the networks optimizing classification accuracy. While genetic (a.k.a. evolutionary) strategies easily incorporate user-defined performance measures and direct optimization, many of the "off the shelf" multivariate classifiers from computer science do not. I can only encourage our field to be more aware of this distinction.

The Bayesian Neural Networks that have been advocated by Radford Neal [24] and used in Ref. [26] preserve a clear connection between the statistical goals of the experiment and the optimization of the multivariate classifier. An **open question** is whether the formalism that he uses provides a practical way to incorporate rate and shape uncertainties in the optimization procedure.

3.2 Matrix-Element Methods

Often, the components of the discriminating variable m are kinematic in nature, eg. masses, momenta, angles, or functions of those quantities. These variables are often strongly and non-trivially correlated, which is why multivariate techniques are so powerful. The kinematic quantities and their correlations are well modeled by a theory. By using Feynman diagrams (a perturbative expansion of Quantum Field Theory) we can readily calculate a complex number called the "matrix element" for an arrangement of initial- and final-state particles. The square of the modulus of the matrix element $|\mathcal{M}|^2$ together with phase space dPS and the parton densities D_{parton} predicts the differential cross section $d\sigma/d\vec{r}$ for the kinematic quantities $\vec{r}$, which is proportional to the probability density function $f(\vec{r})$.

In practice, we use particle-level Monte Carlo programs to sample the distribution $f(\vec{r})$. Since we do not measure the kinematic quantities $\vec{r}$ perfectly, we must also simulate the impact of detector effects, which gives us measured quantities $\vec{r}_m$. The final discriminating quantities m are then calculated from the kinematic quantities $\vec{r}_m$. As previously mentioned, the simulation of the detector can be very complicated and we rely on Monte Carlo techniques for the probabilistic mapping $\vec{r} \rightarrow m$. Standard practice has been to use pseudo-data for m as training data for multivariate classifiers, and, as previously mentioned, the resulting classifiers are often regarded as a "black box".

While the detector simulation is very detailed, the probability that a true $\vec{r}$ results in a measured m can often be approximated with a "transfer function" denoted $W(\vec{r}, m)$. Clearly, a more transparent multivariate approach would be to construct a multivariate classifier by numerically confronting the

convolution of the differential cross-section with the transfer function $W(\vec{r}, m)$.

$$\frac{L(m_j|H_1)}{L(m_j|H_0)} = \frac{\int dPS(\vec{r})\, |\mathcal{M}_{s+b}(\vec{r})|^2\, W(\vec{r}, m_j) D_{parton}(\vec{r})}{\int dPS(\vec{r})\, |\mathcal{M}_{b}(\vec{r})|^2\, W(\vec{r}, m_j) D_{parton}(\vec{r})} \tag{9}$$

The success of this method is limited by the accuracy of the transfer function $W(\vec{r}, m)$ and the computational complexity of the convolution. Modern computers now make the numerical convolution tractable, and experience at the Tevatron shows that these "matrix element techniques" are competitive with other multivariate techniques. These matrix element techniques have been used for the most precise measurements of the top mass [25] and in the context of searches for single top [6] are competitive with boosted decision trees and Bayesian neural networks [26, 27].

In addition to experimental applications of the "matrix element technique", the method has substantial capacity to influence phenomenological studies. A large class of phenomenological studies are sensitivity studies, which ask "what is the sensitivity of a given experiment to a given signature predicted by a new theory?". Traditionally, this question is addressed by generating particle-level Monte Carlo for the kinematic quantities $\vec{r}$, smearing those quantities with a parametrized detector response $W(\vec{r}, \vec{r}_m)$, using creativity and insight to find good discriminating variables $m(\vec{r}_m)$, designing a simple cut-analysis to select signal-like events, and then estimating sensitivity with $s/\sqrt{b}$. In cases where the estimated significance is large, then this theory should be taken seriously and studied in more detail by the experimentalists. However, if the estimated significance is low, it is not clear if the experiment is truly not sensitive to this signature for new physics or if the choice of the discriminating variables and the simple cut analyses were just sub-optimal. To avoid this situation, Tilman Plehn and I considered the use of the matrix element technique as in a phenomenological context [28]. Instead of calculating $L(m_j|H_1)/L(m_j|H_0)$ for an observed event's discriminating variables m_j, one can integrate over the joint distribution $f_{s+b}(\vec{r}_m)$ under the alternate hypothesis and calculate an expected significance. Moreover, since the kinematic variables $\vec{r}_m$ encode all the kinematic information, this expected significance provides an upper-bound. If the upper-bound is low, then one can be sure that the experiment truly is not sensitive to the signature for new physics.

It is worth mentioning that the typical procedure in the matrix element method is to integrate over the "true" particles' kinematics $\vec{r}$. This is comfortable for physicists because that is what we do when we calculate cross-sections and we know the phase-space factors associated with $\vec{r}$. Since we are integrating over "true" quantities, this has a Bayesian feel – but it is a use of Bayes theorem that a Frequentist would not mind because we can consider a frequency distribution for $\vec{r}$. Another point of view might be that for this particular event, the particles had some particular true value, and that it doesn't make sense to talk about a sampling distribution for $\vec{r}$. In that vein, one could imagine $\vec{r}$ to be a vector of nuisance parameters for this event, and that one should choose maximization over integration (marginalization). Such techniques have been considered in the context of supersymmetric mass determination [29]. This point of view brings up several **open questions**: a) which method is more powerful? b) how is $-2\log\lambda$ distributed if new nuisance parameters are added to the problem for each of the x events (which is itself a random variable)? and c) is the maximization approach simpler computationally?

4 Challenges of Searches for Beyond the Standard Model Physics

Sections 2 and 3 considered the scenario in which the signal model was well specified, and focused on the incorporation of the nuisance parameters ν in the background model. Section 2.4 considered the case in which the signal is composite, but the dimensionality of physically significant parameters, γ, in the alternate hypothesis was small enough that the "trials factor" can be readily estimated and no severe loss of power is expected by recalibrating the p-value. In this section, we consider the case where the signal model is composite and γ has many parameters and the case in which the signal is quite vague.

[6] The matrix element techniques used in the D0 searches did not perform as well as the other multivariate techniques; however, it is known that they neglected the matrix element for some of the background processes.

Before continuing, it is worth considering a few specific examples. Perhaps the most studied scenario for "beyond the standard model" physics is supersymmetry (SUSY). If supersymmetry exists, it is a broken symmetry in nature. There is no established mechanism for supersymmetry breaking, so the minimal supersymmetric extension to the Standard Model (MSSM) parametrizes all the soft breaking terms with ~ 105 parameters. Thus, one might look at the unconstrained MSSM as more of a theoretical framework than a theory per se. Within this 105 dimensional space, are a few well-motivated subspaces corresponding to particular scenarios for SUSY breaking, eg. mSUGRA parametrized by four real-valued constants and a sign. Even in this restricted parameter space, the signatures for new physics are quite diverse. In general, one is faced with a generic tradeoff between more powerful searches for specific model points and less-powerful but more robust searches. Despite its complexity, supersymmetric models that conserve something called R-parity, in which there is a generic signature of large missing energy in the detector. This allows for a search strategy that is both powerful (enough) and robust (enough); however, other models do not necessarily have an equivalent generic signature.

There are a host of models in addition to supersymmetry that have been proposed to be relevant to the "terra scale" and accessible to the LHC. To give a feeling for the activity, there have been 32,000 papers in hep-ph since 2000. Clearly, even 4000 physicists cannot give due consideration to all of the proposed models in the landscape. As a result, it is interesting to consider more radical approaches and formalisms that can be applied more generically.

4.1 False Discovery Rate

In 1995, Benjamini and Hochberg introduced a technique called False Discovery Rate (FDR) to confront the challenge of multiple hypothesis tests [30]. In contrast to the Bonferroni-type corrections (eg. trials factor), which seeks to control the chance of even a single false discovery among all the tests performed, the FDR method controls the proportion of errors among those tests whose null hypotheses were rejected. Since that time, FDR has become quite popular in astrophysical data analysis [31]. The properties which make FDR popular are

- It has a higher probability of correctly detecting real deviations between model and data.
- It controls a scientifically relevant quantity: the average fraction of false discoveries over the total number of discoveries.
- Only a trivial adjustment to the basic method is required to handle correlated data.

The definition of the false discovery rate is given by

$$FDR = \frac{N^{\text{reject}}_{\text{null true}}}{N^{\text{reject}}} \tag{10}$$

While it seems like a vacuous re-casting of Type-I and Type-II error, it is fundamentally different since it controls a property of a set of discoveries. In brief the FDR technique is implemented by performing N tests, ordering the tests according to the observed p-values (ascending), specifying an acceptable false discovery rate q, and then rejecting the null hypothesis for the first r tests, where r is the largest j that satisfies $p_j < jq/C_N N$. The quantity C_N is only needed if the tests are correlated. This definition means that the value of the threshold on the largest p-value is not known a priori, but is adaptive to the data set. Furthermore, one does not need to specify an alternate hypothesis – though p-values determined from test statistics that are designed with an alternate in mind can be expected to be more powerful.

FDR has not been widely used within High Energy Physics, but it seems to have a natural place in the context of searches for exotica. While our experiments do not have enough resources to study every model that will be proposed, we do have several clever collaborators and an enormous amount of computing power, so we need to address the multiple testing problem. I do not see FDR as being particularly relevant for searches when we expect to claim only a single discovery (eg. the Higgs); however, I do see that it might have a role in the global analysis of LHC data.

4.2 Interpreting New Physics: The Inverse Problem

If we are lucky enough to find evidence of physics beyond the standard model, the next order of business will be to interpret what we have observed and measure the relevant parameters of the new standard model (sometimes called the "inverse problem" [34] historically just called "physics"). Perhaps we will observe new physics that is consistent with an already well-developed theory; perhaps we will observe something consistent with several different theories and we will need to discriminate between them; or perhaps we will observe something more unexpected and be stumped for some time to provide a concise description or even identify the fundamental parameters to be measured.

4.3 Parameter Scans

For cases such as supersymmetry, in which we have a well-developed theory with known fundamental parameters, it is common to simply scan the parameters on a naive grid. At each parameter point in the scan, one might consider an optimized analysis for that point using an automated procedure [32], or test the consistency of the model with data with the aim to measure the fundamental quantities [33]. Naive parameter scans face two problems, one practical and one conceptual. The practical problems are that grid-scans don't scale to high-dimensions and that simple Monte Carlo sampling is not very efficient. Markov Chain Monte Carlo techniques address the practical problems and are addressed below. The conceptual problem is that the space of the parameters does not have a natural metric – why do we take equal steps in $\tan\beta$ and not in β? Information Geometry provides an elegant, though computationally challenging, solution by equipping the space of the parameters with an experimentally relevant metric. Information Geometry may also provide a useful tool for theorists to formalize the "cliffs" and "valleys" in the landscape.

4.3.1 Markov Chain Monte Carlo & Hierarchical Bayes

Markov Chain Monte Carlo (MCMC) has been used successfully (mainly by the collaborations of theorists and experimentalists) to map out the regions of the constrained MSSM (CMSSM) preferred by existing Standard Model and astrophysical measurements [35, 36, 37, 38]. The CMSSM is described by four parameters and a sign. Typically, the MCMC scans provide a Bayesian posterior distribution for the parameters. As Cousins critiqued in his proceedings to this conference, the groups often use flat priors in relatively high dimensions.

Recently, a similar analysis was performed with a more theoretically driven choice for the prior [39]. There, the authors considered the prior probability density for a given SUSY breaking scale M_S:

$$\begin{aligned} p(m_0, M_{1/2}, A_0, \mu, B, sm|M_S) &= p(m_0|M_S)\, p(M_{1/2}|M_S)\, p(A_0|M_S) \\ &\quad p(\mu|M_S)\, p(B|M_S)\, p(sm), \end{aligned} \tag{11}$$

assuming that the SM experimental inputs sm do not depend upon M_S. A particular choice was made relating the SUSY breaking scale M_S and the parameters m_0, $M_{1/2}$, A_0, by relating them to M_S at the "order of magnitude" level:

$$p(m_0|M_S) = \frac{1}{\sqrt{2\pi w^2} m_0} \exp\left(-\frac{1}{2w^2}\log^2(\frac{m_0}{M_S})\right). \tag{12}$$

The parameters A_0 and B are allowed to have positive or negative signs and values may pass through zero, so a prior of a different form was used:

$$p(A_0|M_S) = \frac{1}{\sqrt{2\pi e^{2w}} M_S} \exp\left(-\frac{1}{2(e^{2w})}\frac{A_0^2}{M_S^2}\right). \tag{13}$$

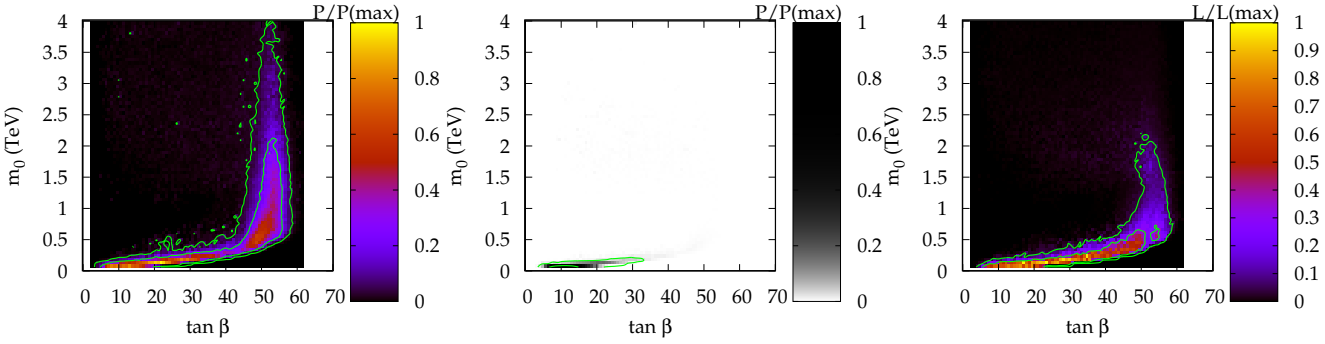

Fig. 1: CMSSM fits marginalised in the unseen dimensions for (a) flat $\tan\beta$ priors, (b) the hierarchical prior with $w = 1$. Figure (c) shows the result of the profile likelihood ratio, in which the unseen dimensions are evaluated at their conditional maximum likelihood values. Contours showing the 68% and 95% regions are shown in each case. The posterior probability in each bin of (a) and (b), normalized to the probability of the maximum bin, is displayed by reference to the color bar on the right hand side of each plot.

Finally, since one does not know M_S a priori, it was treated as a "hyper-parameter" and marginalized giving

$$p(m_0, M_{1/2}, A_0, \mu, B) = \int_0^\infty dM_S\, p(m_0, M_{1/2}, A_0, \mu, B|M_S)\, p(M_S) \tag{14}$$

$$= \frac{1}{(2\pi)^{5/2} w^5 m_0 |\mu| M_{1/2}} \int_0^\infty \frac{dM_S}{M_S^2} \exp\left[-\frac{1}{2w^2}\left(\log^2(\frac{m_0}{M_S}) + \log^2(\frac{|\mu|}{M_S}) + \right.\right.$$
$$\left.\left. \log^2(\frac{M_{1/2}}{M_S}) + \frac{w^2 A_0^2}{e^{2w} M_S^2} + \frac{w^2 B^2}{M_S^2 e^{2w}}\right)\right] p(M_S),$$

where $p(M_S)$ is the prior for M_S itself, which was taken to be flat in the logarithm of M_S. The marginalisation over M_S amounts to a marginalisation over a family of prior distributions, and as such constitutes a hierarchical Bayesian approach. Fig. 1 shows a comparison between the results obtained with flat priors (a) and those obtained with the hierarchical approach (b). As far as I am aware, Ref. [39] is the first example of the use of hierarchical Bayesian techniques in particle physics.

4.3.2 Frequentist Approach

It is clear from Fig. 1 that the choice of prior has a large effect on the results obtained. In the sense of "forecasting" what the LHC might see, the hierarchical approach is playing an important role by injecting our physical insight and sharpening our focus. However, in terms of an experimental result the dependence on a prior is often seen as undesirable, thus it is interesting to consider frequentist approaches.

The MCMC scans were performed in a four-dimensional parameter space, but the figures show two-dimensional projections. In the Bayesian approach, one marginalizes the unseen dimensions with respect to the prior. A frequentist analysis would eliminate the unseen dimensions by maximization instead of marginalization – eg. use the profile likelihood ratio. Fig. 1(c) shows the result of the same analysis with the profile likelihood ratio. We see similar constraints, except that the tail at high $\tan\beta$ up to larger values of $m_0 > 2$ TeV has been suppressed in the profile. From the difference we learn the following facts: in this high $\tan\beta$-high m_0 tail, the fit to data is less good than in other regions of parameter space. However, it has a relatively large volume in unseen dimensions of parameter space, which enhances the posterior probability in Fig. 1(a). The difference between the two plots is therefore a good measure of the so-called "volume effect". While one may argue that flat priors distort the inference by pushing all the probability away from the origin, it is clear that the hierarchical priors had much more of an effect on the inference (reflecting the fact that the data are not dominating the Bayesian inference).

Other groups have performed frequentist analyses of essentially the same problem, though without the use of MCMC to scan the parameter space [40, 41]. In both cases the asymptotic distribution of the profile likelihood ratio was used in constructing confidence intervals. Given the complexity of the likelihood function, it is an **open question** if the asymptotic χ^2 distributions provide good coverage properties for these studies.

4.4 Information Geometry

Information Geometry is a synthesis of statistics and differential geometry. In essence information geometry equips model space with a "natural" metric that is invariant to reparametrization of observables, m, and covariant to reparametrization of theoretical parameters, γ [42].

$$g_{ij}(\gamma) = \int dm f(m;\gamma) \left[\frac{\partial \log f(m;\gamma)}{\partial \gamma_i}\right] \left[\frac{\partial \log f(m;\gamma)}{\partial \gamma_j}\right] \tag{15}$$

By equipping the space of the models with a metric, one can do many powerful things. It has been shown in the context of machine learning that learning algorithms that take equal steps in this natural geometry can converge exponentially faster than one that takes equal steps in the naive parameters of the learning machine. In the context of experimental high energy physics, one can imagine that Information Geometry could make parameter scans significantly more efficient.

Information Geometry may play an even more useful role in theoretical analyses. For instance, the authors of Ref. [34] considered a 15-dimensional supersymmetric model and an exhaustive list of relevant observables. The authors sought to analyze the structure of this space by finding degeneracies (ie. points γ_a and γ_b where the observables are essentially unchanged) and "cliffs" (ie. regions where a small change in γ gives rise to a large change in the observables). These questions could be addressed formally if one had access to the metric $g_{ij}(\gamma)$. Instead, their analysis used a rather ad hoc $\Delta\chi^2$-like discriminant for the observables and a non-invariant Euclidean-like distance for the parameter space γ. While their results seemed quite reasonable, and the degeneracies they found correspond to physically reasonable scenarios, it would be a significant advance if such studies could be formalized.

4.5 Interpretation and The Theory-Experiment Interface

Another challenge of beyond the standard model searches is how to represent the result of an observation and communicate sufficient information to the field. Because of the complexity of some of the models it is not possible to represent the results as a simple one-dimensional likelihood curve or a two-dimensional contour without substantial loss of information. For instance, Fig. 1 only shows two of four interesting dimensions in the theory's parameter space. Ideally, experiments would publish a likelihood map in the full dimensionality of relevant quantities – this is technically possible in many cases [39] and a new feature of the RooFit/RooStats framework [43].

Another issue for publication is model-dependence; it is common for a single experimental signature to be described by several models with different fundamental parameters. It is not feasible for the experiments to report the results tailored for each conceivable model. Instead, experiments prefer to report their results in a model-independent way. In some cases (eg. different models for a Z') there is a model-relevant and model-neutral set of parameters that can be measured, which encompass several different theoretical models, while still providing enough information to distinguish among them. This is an ideal case, but it is not always obvious which measurements are sufficient to distinguish between competing models.

Recently an old theoretical tool (on-shell effective theory) was given a new spin, in the form of a toolset called MARMOSET [44]. MARMOSET is meant to quickly provide a simplified description of new physics, especially in cases where the data are not described by an already well-developed theoretical model. Despite its simplifications, the authors of MARMOSET argue that it does maintain the

essential features of many scenarios for new physics. It is an **open question** if full likelihood maps of the parameters of the best fitting on-shell effective theories provide a general purpose solution for model-neutral and model-relevant publications for the LHC.

5 Conclusion

We are entering a very exciting time for particle physics. The LHC will be probing the "tera-scale", which may reveal the mechanism for electroweak symmetry breaking, new symmetries of nature, and evidence for additional space-time dimensions. The rich landscape of theoretical possibilities and the particularly challenging experimental environment of the LHC place particular emphasis on our statistical techniques. Searches for specific signatures, like the Higgs boson, must address large background uncertainties and consistently combine several search channels. Substantial progress has been made in terms of incorporating systematics in our statistical machinery. Searches for beyond standard model physics have additional challenges, which are more strategical in nature. In particular, how should we approach the search when the signal model is vague or the model space is very large and the phenomenology is diverse? We still have not fully addressed the multiple testing problem for the LHC, but perhaps methods like False Discovery Rate have a role to play in the global analysis of the LHC data. If we are fortunate enough to discover new physics at the LHC, we will begin the process of interpreting what we saw. Perhaps we will see something expected and the process will be fairly straightforward; however, we must be prepared for something more unexpected. Ideally, the experiments will publish their results in terms of a full likelihood scan of a model-neutral and model-relevant parameter space. The technical challenge of reporting a full likelihood map has been addressed by the RooFit framework, the remaining challenge is choosing how to represent the data. In some cases the field has already converged on an appropriate set of parameters to measure for a given signature, but we do not have an adequate solution in the case of something more unexpected. A recent proposal is to use on-shell effective theories as a concise summary of LHC data, which could provide a general purpose solution for publishing model-neutral and model-relevant results. While there remain many open questions to address, the PhyStat conference series has been very effective in preparing our field for the statistical challenges of the LHC.

References

[1] K. Cranmer, *proceedings of PhyStat05, Oxford, England, United Kingdom, (2005)*

[2] S. Frixione and B. R. Webber, JHEP **0206** (2002) 029 [arXiv:hep-ph/0204244].

[3] Andreas Schalicke and Frank Krauss. *JHEP*, 07:018, (2005).

[4] LEP Higgs Working Group. *Phys. Lett.*, B565:61–75, (2003).

[5] Nancy Reid. proceedings of PhyStat2003, (2003).

[6] W. A. Rolke and A. M. Lopez. *Nucl. Instrum. Meth.*, A458:745–758, (2001).

[7] W. A. Rolke, A. M. Lopez, and J. Conrad. *Nucl. Instrum. Meth.*, A551:493–503, (2005).

[8] K. Cranmer. *proceedings of PhyStat2003* (2003) [physics/0310108].

[9] G. Punzi. *proceedings of PhyStat2005* (2005) [physics/0511202].

[10] R.D. Cousins and V.L. Highland. *Nucl. Instrum. Meth.*, A320:331–335, (1992).

[11] J. Linnemann. *proceedings of PhyStat2003* [physics/0312059], (2003).

[12] R.D. Cousins and J. Tucker, [arXiv:physics/0702156] (2007)

[13] F. James and M. Roos. *Comput.Phys.Commun.*, 10:343–367, (1975).

[14] S. Asai et al. *Eur. Phys. J.*, C3252:19–54, (2004).

[15] L. Demortier p-values *these proceedings*, PhyStat2007 (2007).

[16] W. Quayle *talk presented at*, PhyStat2007 (2007).

[17] K. Cranmer, "LEPStats4LHC", software available at *phystat.org* repository. [packages/0703002].

[18] W. A. Rolke and A. M. Lopez, [arXiv:physics/0606006].

[19] K. Cranmer, *In the Proceedings of PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology, WEJT002 (2003),* [arXiv:physics/0310110].

[20] K. Cranmer. *Acta Phys. Polon.*, B34:6049–6068, (2003).

[21] K. Cranmer and R. S. Bowman. *Comp. Phys. Commun.*, 167(3):165–176, (2005).

[22] J. M. Link *et al.* [FOCUS Collaboration], Phys. Lett. B **624** (2005) 166 [arXiv:hep-ex/0507103].

[23] S. Whiteson and D. Whiteson, [arXiv:hep-ex/0607012].

[24] R.M. Neal, *Bayesian Learning of Neural Networks*. Springer-Verlag, New York, (1996)

[25] V. M. Abazov *et al.* [D0 Collaboration], Nature **429** (2004) 638 [arXiv:hep-ex/0406031].

[26] V. M. Abazov *et al.* [D0 Collaboration], Phys. Rev. Lett. **98** (2007) 181802 [arXiv:hep-ex/0612052].

[27] W. Wagner, [hep-ex/0610074]; The CDF Collaboration, public conference note 8185, April 2006; M. Bühler, Diplomarbeit Universität Karlsruhe, FERMILAB-MASTERS-2006-02, (2006).

[28] K. Cranmer and T. Plehn, Eur. Phys. J. C **51**, 415 (2007). [arXiv:hep-ph/0605268].

[29] C.G. Lester, "Part X:" , *Les Houches 'Physics at TeV Colliders 2003' Beyond the Standard Model Working Group: Summary report,* [arXiv:hep-ph/0402295].

[30] Y. Benjamini and Y. Hochberg *J. R. Stat. Soc.*-B 57:289-300 (1995)

[31] C. J. Miller *et al.*, Astro.Jour., 122.6,3492-3505 (2001) [arXiv:astro-ph/0107034].

[32] V. M. Abazov *et al.* [D0 Collaboration], Phys. Rev. Lett. **87**, 231801 (2001)

[33] I. Hinchliffe, F. E. Paige, M. D. Shapiro, J. Soderqvist, and W. Yao. *Phys. Rev.*, D55, (1997).

[34] N. Arkani-Hamed, G. L. Kane, J. Thaler and L. T. Wang, JHEP **0608** (2006) 070

[35] E. A. Baltz and P. Gondolo, JHEP **0410** (2004) 052 [arXiv:hep-ph/0407039].

[36] B. C. Allanach and C. G. Lester, Phys. Rev. D **73** (2006) 015013 [arXiv:hep-ph/0507283].

[37] R. R. de Austri, R. Trotta and L. Roszkowski, JHEP **0605** (2006) 002 [arXiv:hep-ph/0602028].

[38] M. Rauch, R. Lafaye, T. Plehn and D. Zerwas, arXiv:0710.2822 [hep-ph].

[39] B. C. Allanach, K. Cranmer, C. G. Lester and A. M. Weber, JHEP **08**, 023 (2007). Likelihood maps published at *http://users.hepforge.org/∼allanach/benchmarks/kismet.html*

[40] W. de Boer, M. Huber, C. Sander and D. I. Kazakov, Phys. Lett. B **515** (2001) 283.

[41] J. R. Ellis, K. A. Olive, Y. Santoso and V. C. Spanos, Phys. Rev. D **69** (2004) 095004

[42] S. Amari, *Differential-geometrical methods in statistics*, Springer-Verlag, Berlin, 1985

[43] W. Verkerke, *http://roofit.sourceforge.net/* and *these proceedings*

[44] N. Arkani-Hamed, P. Schuster, N. Toro, J. Thaler, L. T. Wang, B. Knuteson and S. Mrenna, arXiv:hep-ph/0703088.