PROCEEDINGS OF SCIENCE

Statistical combination of charged Higgs channels in ATLAS

Ofer Vitells[1]
Weizmann Institute of Science
Rehovot, Israel
E-mail: ofer.vitells@weizmann.ac.il

Eilam Gross
Weizmann Institute of Science
Rehovot, Israel
E-mail: eilam.gross@weizmann.ac.il

We describe the statistical procedures for combination of results from independent searches for the charged Higgs boson. The methods are applied to Monte Carlo studies of five search channels currently considered at the ATLAS experiment. The statistical treatment is based on a frequentist approach, where effects of systematic uncertainties are incorporated by use of the profile likelihood ratio. Results are presented for the expected statistical significance of discovery and expected exclusion limits.

Prospects for Charged Higgs Discovery at Colliders
Uppsala, Sweden
September 16th-19th 2008

[1] Speaker

PoS(CHARGED2008)038

1.Introduction

Charged Higgs searches will exploit a number of statistically independent channels. One wishes to combine all of the information from them to provide a single measure of the significance of a discovery or limits on charged Higgs production. In this note we describe the statistical methods used for the combination of the five search channels currently studied with ATLAS simulation [1].The approach taken here is frequentist based, where effects of systematic and statistical uncertainties are incorporated by use of the profile likelihood ratio [2].

The use of the likelihood ratio as a test statistics that summarizes the data is fairly common, see for example [3]. In particular, the sampling distribution of the likelihood ratio is known in the large sample limit [4]. This allows for a fast evaluation of discovery and exclusion limit over a wide range of signal parameters, which is needed if one wants to establish sensitivity regions within theoretical models such as the MSSM.

The profile likelihood method is described in section 2. In section 3 we describe the specific statistical model used for the charged Higgs combination, and in section 4 we show the results of the combination and their interpretation within MSSM scenarios.

2.The Profile Likelihood method

The profile likelihood method uses a likelihood ratio to distinguish between two hypothses, the background-only hypothesis (H_0) and the signal+background hypothesis (H_1). We introduce a signal normalization parameter μ, such that the expected number of events in a given channel (or equivalently in a single bin of a histogram) is given by:

$$E[n_i] \equiv \mu s_i + b_i \tag{1}$$

i.e., $\mu = 1\,(0)$ corresponds to the $H_1\,(H_0)$ hypothesis. The expected numbers of signal and background events, s_i and b_i, may not be known precisely a-priori, but rather estimated with some theoretical and experimental uncertainties. In such case they are called *nuisance parameters*. More generally, we can have a set of nuisance parameters $\boldsymbol{\theta}$ related to the expected signal and background rates (such as luminosity, variuos experimental efficiencies etc.). We then define a likelihood function that describes the probability of observing n_i events for a given values of the nuisance parameters:

$$L(n_i \mid \mu, \boldsymbol{\theta}) \tag{2}$$

The *profile likelihood* ratio, $\lambda(\mu)$, is defined as:

$$\lambda(\mu) = \frac{L(n_i \mid \mu, \hat{\hat{\boldsymbol{\theta}}})}{L(n_i \mid \hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{3}$$

where $\hat{\mu}, \hat{\boldsymbol{\theta}}$ are the maximum likelihood estimators (MLE) of $\mu, \boldsymbol{\theta}$ respectively, and $\hat{\hat{\boldsymbol{\theta}}}$ is the conditional MLE of $\boldsymbol{\theta}$ for a fixed value of μ. The likelihood ratio is defined such that $0 \le \lambda(\mu) \le 1$. Wilks' theorem [4] states that the sampling distribution of $-2\log\lambda(\mu)$ approaches a χ^2 distribution with one degree of freedom, when the data n_i is consistent with μ. For example, this is true for $-2\log\lambda(\mu=0)$ when the data is consistent with the background-only hypothesis, and for $-2\log\lambda(\mu=1)$ when the data is consistent with the signal+background hypothesis.

2.1 Statistical significance of a discovery

The statistical significance of observing a signal is usually defined in terms of the rate of 'Type-I errors'. That is, the probablity ('p-value') that a background fluctuation will fake a signal. In terms of the profile likelihood ratio (3), this can be expressed as the probabilty of $-2\log\lambda(\mu=0)$ being larger then it's observed value, under the background-only hypothesis. This is schematically illustrated in Figure 1.

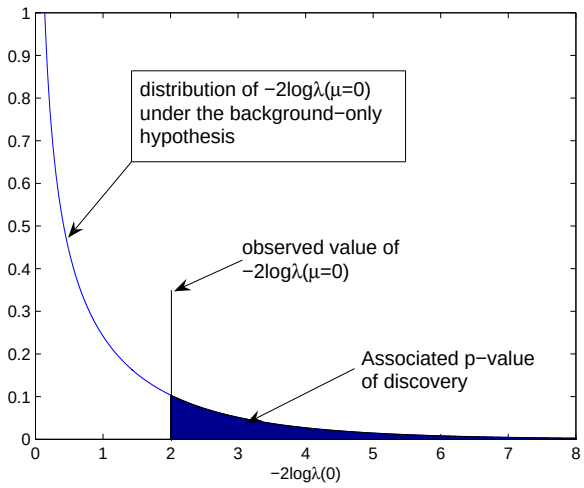

Figure 1 Illustration of the relation between the observed value of $-2\log\lambda(\mu=0)$ *and the p-value. The shaded area represents the probablity that a background fluctuation will give rise to value of* $-2\log\lambda$ *that is equal to or larger then the one actually observed.*

Using the properties of the χ^2 distribution, one can express the associated p-value in a more common form as a number of standard deviations, using the relation:

$$Z_{discovery} = \sqrt{-2\log\lambda(\mu=0)} \tag{4}$$

2.2 Exclusion limits

The statistical significance of excluding a signal can be similarly derived from the expected distribution of the profile likelihood ratio. In this case the situation is reversed and one is interested in the probability of rejecting the signal+background hypothesis when it is actually true. The significance in terms of standerd deviations is similarly given by:

$$Z_{exclusion} = \sqrt{-2\log\lambda(\mu=1)} \tag{5}$$

2.3 Combined significance

In the case of several independent search channels, the method described above is generalized in a straight forward manner. The full likelihood function is simply given by the product of the individual ones:

$$L(n \mid \mu) = \prod_i L_i(n_i \mid \mu, \boldsymbol{\theta}_i) \tag{6}$$

In principle some of nuisance parameters $\boldsymbol{\theta}_i$ may be common to several channels, however we keep the general notation for simplicity. The construction of the profile likelihood ratio (3) now requires a simultaneous maximization of the likelihood function with respect to all of the nuisance parameters as well as the signal strength parameter μ.

The situation is simplified when one wishes to estimate *expected* sensitivity based on a Monte Carlo simulation. In this case one uses $n_i = s_i + b_i$ as the 'observed' number of events in a given channel (under the signal+background hypothesis). That is, the observed numbers are assumed to be equal to their expected values[1]. This is sometimes refferd to as the 'Asimov' dataset [2]. This lead to $\hat{\mu}$ being exactly equal to one both for the global likelihood function and each of the individual ones, $\hat{\mu} = \hat{\mu}_i = 1$, such that the profile likelihood ratio can be written as

$$-2\log\lambda(\mu=0) = \sum_i \log\frac{L(n_i \mid \mu, \hat{\hat{\boldsymbol{\theta}}}_i)}{L(n_i \mid \hat{\mu}, \hat{\boldsymbol{\theta}}_i)} = \sum_i \log\frac{L(n_i \mid \mu, \hat{\hat{\boldsymbol{\theta}}}_i)}{L(n_i \mid \hat{\mu}_i, \hat{\boldsymbol{\theta}}_i)} \tag{7}$$

The combined significance is therefore simply given by the sum of each individual channel's significance, taken in quadrature:

$$Z_{discovery} = \sqrt{Z_1^2 + Z_2^2 + \ldots + Z_n^2} \tag{8}$$

where Z_i denotes the individual discovery significance of channel i. Similar relation holds for the combined significance of exclusion.

3.The statistical model

For the combination of the charged Higgs channels, a simple statistical model of the data is used in which the background uncertainty is taken to be 10% for all channels[3]. The model

[1] In principle, one can generate a set of random numbers distributed around the expected values with Poisson probabilities, so as to mimic a realistic dataset. In that case the resulting significance can be viewed as a random variable who's expected value is given by the 'Asimov' dataset.

[2] Inspired by the short story *Franchise* by Isaac Asimov, in which elections are held by selecting a single voter to represent the entire electorate.

[3] This value is estimated to be the uncertainty on background rates when they are evaluated from data, using specialized methods [5].

also takes into account the statistical uncertainties related to the limited sample size of the Monte Carlo simulation used to evaluate the backgrounds.

Consider a channel in which the number of background events estimated by Monte Carlo is b^{MC}, for some integrated luminosity $\mathcal{L}^{MC}$. We then define the following likelihood function of the number of observed events n, at an arbitrary luminosity $\mathcal{L}$ where $\tau \equiv \mathcal{L} / \mathcal{L}^{MC}$

$$L(n \mid \mu) = Poiss(n \mid \mu s + \tau b) Poiss(b^{MC} \mid b_0^{MC}) Gamma(b_0^{MC} \mid b, \sigma_b) \quad (9)$$

Here b_0^{MC} is the expected number of Monte Carlo events, which is associated with a systematic uncertainty σ_b. b is the true number of expected events at an integrated luminosity $\mathcal{L}^{MC}$. The choice of a Gamma distribution for representing the systematic uncertainty is of course somewhat arbitrary, since a systematic uncertainty by nature cannot be associated with a distribution in the usual probabilistic sense.

The likelihood function (9) does not include any uncertainties related to the number of signal events s. For discovery significance such uncertainties are not relevant since the profile likelihood ratio is evaluated at $\mu = 0$, however they are important for exclusion[4]. In that case similar factors are added to (9) with respect to s.

Finally, a slight complication is caused by the fact that some of the backgrounds used for the analysis are estimated by event generators that produce events with positive and negative weights. These need to be considered separately since they both contribute to the statistical uncertainty related to the total number of backgrounds events. The likelihood function (9) is modified in this case to:

$$L(n \mid \mu) = Poiss(n \mid \mu s + \tau b) Poiss(b_{pos}^{MC} \mid b_0^{pos}) Poiss(b_{neg}^{MC} \mid b_0^{neg}) Gamma(b_0^{pos} - b_0^{neg} \mid b, \sigma_b) \quad (10)$$

where $b_{pos}^{MC}, b_{neg}^{MC}$ are the numbers of positive and negative Monte Carlo events, and b_0^{pos}, b_0^{neg} are their expected values, respectively.

4.Results

The expected significance for both discovery and exclusion are evaluated as a function of the charged Higgs production cross section. In Figure 2 the contours corresponding to a discovery significance of 5σ and exclusion at a 95% confidence level are shown in the $(m_{H^+}, \tan\beta)$ plane within the MSSM $m_h - \max$ scenario, for an integrated luminosities of 1, 10 and 30 fb^{-1}.

For the analyses considering a light charged Higgs $(m_{H^+} < m_{top})$ the Monte Carlo background sample size was approximately equivalent to 1 fb^{-1} of data. The expected number of background events at a luminosity of 30 fb^{-1} is therefore associated with a large statistical uncertainty, which effectively does not allow extending the discovery sensitivity reach. This effect is clearly seen in Figure 2. In order to assess how the situation *might* be improved with a

[4] This is easily understood: a discovery can be made if a significant excess of events above the background is observed, even if one does not know in advance what the signal rate is. On the other hand, a signal cannot be excluded if It's value is not known.

larger sample size, we show in Figure 3 the same contours in the case that the background estimates are assumed to be based on an infinitely large Monte Carlo sample. It should be clear however that precise determination of the expected sensitivity at luminosities higher then 1 fb^{-1} would only be possible if a sufficient amount of simulated events is available.

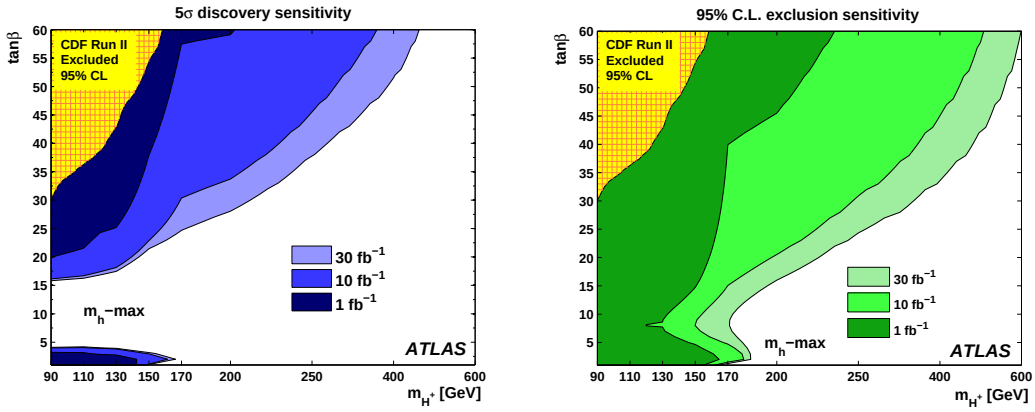

Figure 2. combined discovery(right) and exclusion(left) contours in the MSSM m_h-max scenario, for an integrated luminosities of 1,10 and 30 fb^{-1}.

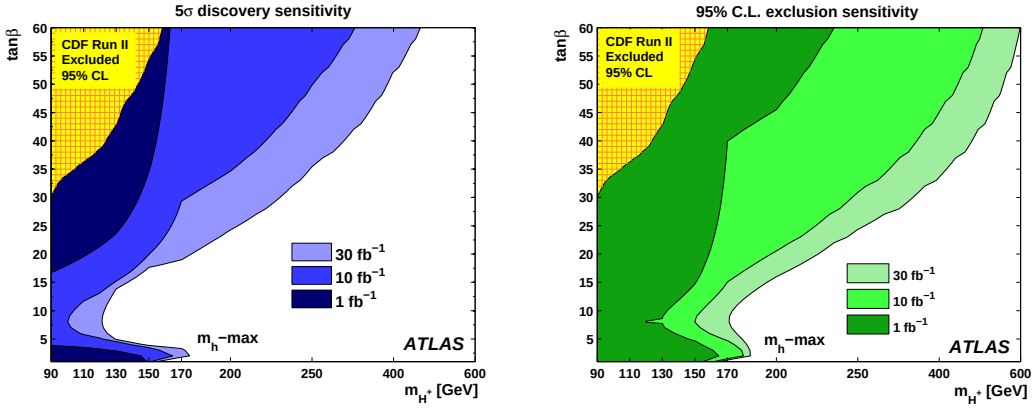

Figure 3. combined discovery(right) and exclusion(left) contours when statistical uncertainties related to the Monte Carlo sample size are ignored.

Conclusions

The procedure for combination of search results based on the profile likelihood ratio has been applied to a study of the search for the charged Higgs boson. The profile likelihood method allows one to construct a test statistic that naturally takes into account systematic uncertainties as well as Monte Carlo statistics used to estimate backgrounds. The asymptotic approximation of the sampling distribution of the likelihood ratio allows for a fast evaluation of discovery and exclusion sensitivities in various theoretical scenarios.

The expected discovery sensitivity with a 5σ significance as well as the exclusion sensitivity at 95% confidence level have been obtained as a function of the charged Higgs mass

and $\tan\beta$ for the m_h-max scenario within the MSSM. For current analyses, sensitivity estimates at high luminosities are limited by available Monte Carlo statistics.

References

[1] M. Flechl , *Charged Higgs prospects with ATLAS*, This volume.

[2] M.S. Bartlett, Biometrica **40** (1953) 306; D.N. Lawley, Biometrica **43** (1956) 295; S.A. Murphy, A.W. Van Der vaart, J. Am. Statist. Assoc. **95** (2000) 449.

[3] N. Reid, D.A.S. Fraser, *Likelihood Inference in the Presence of Nuisance Parameters,* In the proceedings of PHYSTAT2003,Menlo park, California,8-11 Sep 2003, pp THAT001; J. Conrad, F. Tegenfeldt, *Likelihood Ratio Confidence Intervals with Bayesian Treatment of Systematic Uncertainties*, In the proceedings of PHYSTAT05, Oxford, England, 12-15 Sep 2005.

[4] Wilks, S. S. *The large-sample distribution of the likelihood ratio for testing composite hypotheses*. Ann. Math. Statist. **9** (1938) 60.

[5] T. Vickey, *Data-driven methods for the estimation of ttbar backgrounds to charged Higgs searches*, This volume.

[6] I. Asimov, *Franchise*, in *Isacc Asimov: The Complete Stories, Vol. 1*, Broadway books, 1990.