Glen Cowan
DRAFT
4 September, 2007

Bayesian methods for the ATLAS Higgs Search

The following note discusses Bayesian statistical methods for the ATLAS Higgs search. It follows a previous (draft) note [1] which described frequentist procedures. Section 1 covers general aspects of Bayesian statistics, while Section 2 applies the methods to to the Higgs search.

1 General framework of Bayesian statistics

In this section the main components of a Bayesian analysis are summarized. Section 1.1 covers subjective probability, Bayes' theorem and marginalization. In Section 1.2 we describe Markov Chain Monte Carlo, which provides an important computational tool. Setting limits is discussed in Section 1.3 and hypothesis testing (i.e., model selection or 'discovery') is treated in Section 1.4. Bayesian hypothesis testing or involves a quantity called the Bayes factor. Obtaining these entails a number of computational challenges that are considered in Section 1.5.

Useful books on Bayesian methods include the texts by Ghosh et al. [2], O'Hagan [3] and Gregory [4].

1.1 Subjective probability, Bayes' theorem, marginalization

In frequentist statistics, probability is associated only with data, i.e., with the outcome of a repeatable observation. Its value is interpreted as the long term frequency with which the outcome should occur. In Bayesian statistics, the meaning of probability is extended to include degree of belief, and thus one can speak of the probability that a given model is true or that the true value of a parameter lies within a certain fixed range.

Let $\mathbf{x}$ represent the data outcome of an observation and assume we have a model that predicts the probability for $\mathbf{x}$, $L(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is set of model parameters whose values are uncertain. When regarded as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{x}$, this is simply the likelihood function.

In Bayesian statistics one uses the likelihood as the conditional probability for $\mathbf{x}$ given $\boldsymbol{\theta}$. Bayes' theorem relates this to the conditional probability for $\boldsymbol{\theta}$ given $\mathbf{x}$ as

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{P(\mathbf{x})} = \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,d\boldsymbol{\theta}'} \ . \tag{1}$$

Here $\pi(\boldsymbol{\theta})$ is the prior probability for $\boldsymbol{\theta}$, or simply 'the prior'. It represents one's degree of belief about the value of $\boldsymbol{\theta}$ before seeing the outcome of the experiment.

The posterior probability $p(\boldsymbol{\theta}|\mathbf{x})$ encapsulates all of one's knowledge about $\boldsymbol{\theta}$ given the observation $\mathbf{x}$ and the prior probabilities $\pi(\boldsymbol{\theta})$. The prior probability for the data $P(\mathbf{x})$ plays the role of the normalization constant for the posterior pdf.

Bayesian statistics provides no unique recipe for writing down the prior $\pi(\boldsymbol{\theta})$, which could be based on previous measurements, theoretical prejudices or other subjective judgements.

Once the prior has been specified, however, Bayes' theorem determines uniquely how the probability of the hypothesized $\boldsymbol{\theta}$ should be updated in the light of the data $\mathbf{x}$. In order for the final result to be of value to the broader scientific community, whose members may or may not share the prior beliefs of the analyst, it is important to show how the posterior probabilities change under a reasonable variation of the prior.

Often the set of parameters $\boldsymbol{\theta}$ contains some components $\boldsymbol{\psi}$ that are of interest in the analysis and others $\boldsymbol{\lambda}$ that are not (nuisance parameters). To find the marginal distribution of those parameters of interest, we simply integrate (marginalize) over the unwanted ones, i.e.,

$$p(\boldsymbol{\psi}|\mathbf{x}) = \int p(\boldsymbol{\psi}, \boldsymbol{\lambda}|\mathbf{x}) \, d\boldsymbol{\lambda} . \tag{2}$$

1.2 Bayesian computation with Markov Chain Monte Carlo (MCMC)

In order to determine the marginal density of a parameter of interest, we need to integrate over all of the nuisance parameters, as indicated in equation (2). For some special cases the integrals can be done in closed form. Formulae relevant to counting experiments are can be found in [5, 6, 7].

For more general situations one needs to integrate numerically. If the number of dimensions is small, the integrals of this type can be done in closed form by standard numerical methods such as Gaussian quadrature or by Monte Carlo, e.g., with the acceptance-rejection algorithm. Here one samples values according to the full joint pdf and records the distribution of the parameter of interest only. With an increasing number of dimensions, however, algorithms such as acceptance-rejection will accept a smaller fraction of the generated points, and the method becomes impractical. Calculation of marginal posterior densities using importance sampling MC is investigated in [7].

Recently it has become possible to treat higher dimensional problems using Markov Chain Monte Carlo methods (MCMC). In depth treatments of MCMC can be found, e.g., in the texts by Robert and Casella [8] and Liu [9]. MCMC generates a correlated sequence of points from a pdf $p(\boldsymbol{\theta})$. The fact that the points are not statistically independent is not important for purposes of determining a marginal distribution, although it does mean that the statistical errors do not decrease as rapidly as the naive $\sqrt{n}$ rate that one has with usual MC.

In many applications the Metropolis-Hastings algorithm is the most useful MCMC method. It allows one to generate multidimensional points $\boldsymbol{\theta}$ distributed according to a target pdf that is proportional to a given function $p(\boldsymbol{\theta})$. It is not necessary to have $p(\boldsymbol{\theta})$ normalized to unit area. This is a very useful property since from Bayes' theorem we have $p(\boldsymbol{\theta}|\mathbf{x}) \propto L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where the constant of proportionality is often not known.

To generate points that follow $p(\boldsymbol{\theta})$, we first need another pdf called the proposal density $q(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$. In principle this can be (almost) any pdf from which one can easily generate independent values, e.g., with standard Monte Carlo methods. The proposal density for $\boldsymbol{\theta}$ contains as a parameter another point in the same space $\boldsymbol{\theta}_0$. For example one can use a multivariate Gaussian centred about $\boldsymbol{\theta}_0$. Beginning at an arbitrary starting point $\boldsymbol{\theta}_0$, the Hastings algorithm iterates the following steps:

1. Generate a value $\boldsymbol{\theta}$ using the proposal density $q(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$.

2. Form the Hastings test ratio, $\alpha = \min\left[1, \frac{p(\boldsymbol{\theta})q(\boldsymbol{\theta}_0; \boldsymbol{\theta})}{p(\boldsymbol{\theta}_0)q(\boldsymbol{\theta}; \boldsymbol{\theta}_0)}\right]$.

3. Generate a value u uniformly distributed in $[0, 1]$.

4. If $u \leq \alpha$, move to the proposed point, i.e., $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$. Otherwise, repeat the old point, i.e., $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$.

Often one takes the proposal density to be symmetric in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. This is then called the Metropolis-Hastings algorithm, and the test ratio becomes

$$\alpha = \min\left[1, \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_0)}\right] . \tag{3}$$

That is, if the proposed $\boldsymbol{\theta}$ is at a value of probability higher than $\boldsymbol{\theta}_0$, the step is taken. If the proposed step is rejected, hop in place.

One can show that the sequence of generated points will populate the space according to the target density only in the limit that it runs forever. For a finite sequence there may be an initial 'burn-in' period during which the points do not follow $p(\boldsymbol{\theta})$. This initial part of the sequence would be excluded from any subsequent calculations. If one chooses the initial point to be in a region of reasonably high probability, however, then it has been argued that no burn-in period is needed. Unfortunately there do not exist useful theorems that might guarantee that the burn-in stage has passed or that one has generated enough points so as to be sure one has adequately sampled the entire space.

One way of assessing the performance of the algorithm is to compute the autocorrelation as a function of the lag k. This is defined as

$$\rho_k = \frac{\text{cov}[\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+k}]}{V[\boldsymbol{\theta}_i]} , \tag{4}$$

where $\text{cov}[\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+k}]$ is the covariance of a generated point with the one k steps away, and $V[\boldsymbol{\theta}_i]$ is the variance of the generated points. One would like the autocorrelation to fall to zero as quickly as possible with increasing lag.

Generally one chooses the proposal density so as to optimize the algorithm according to, say, the autocorrelation. For certain problems it has been shown that one achieves optimal performance when the acceptance fraction, that is, the fraction of points with $u \leq \alpha$, is around 40%. This can be adjusted by varying the width of the proposal density. This can be done by using a multivariate Gaussian with a covariance matrix U aligned to that of the target pdf V, but scaled by a constant, i.e.,

$$U = cV . \tag{5}$$

The covariance matrix V of the target pdf can be found from the matrix of second derivatives evaluated at the target's mode, e.g., using a program such as MINUIT. The constant c is adjusted to give an acceptance fraction close to 0.4. Often values in the range $0.5 < c < 2$ are found.

In some problems one may find that the Gaussian tails of the proposal density fall off so quickly that one cannot reach regions of $p(\boldsymbol{\theta})$ that may be separated by valleys of low probability. In such cases the performance may be improved by using a proposal density with longer tails, such as a multivariate Student's t distribution.

1.3 Bayesian limits

Limits or intervals for parameters can be constructed by integrating the posterior probability so as to include any desired probability. For a single parameter θ, for example, an upper limit at confidence level $1 - \alpha$ is determined from

$$1 - \alpha = \int_{-\infty}^{\theta_{\text{up}}} p(\theta|\mathbf{x})\, d\theta \ . \tag{6}$$

Often one would choose $1 - \alpha$ to be some conventional value such as 0.9 or 0.95. If the parameter is known to be non-negative (e.g., a cross section), then the prior, and thus also the posterior, would be zero for $\theta < 0$, and the integral (6) effectively starts at a lower limit of zero.

Intervals constructed via equation (6) are often called Bayesian credible intervals, to distinguish them from frequentist confidence intervals. The latter are constructed so as to cover the true value of a parameter with a specified probability (the confidence level) independent of the parameter's true value. That is, the coverage is the fraction of times that an interval constructed according to a given prescription from the data would contain the parameter's value after many repetitions of the experiment. The coverage probability of a Bayesian credible interval could in general be greater or less than its probability content $1 - \alpha$, and it is often of interest to investigate, e.g., with Monte Carlo experiments, the coverage as a function of θ for at least some range of values.

1.4 Bayesian discovery (model selection)

In frequentist statistics, a discovery would be established by constructing a test of the null (no signal) hypothesis H_0 with a high power relative to an alternative H_1 (e.g., Higgs at a certain mass). One finds a p-value of H_0, which is the probability to see data, under the assumption of H_0, which is at least as incompatible with H_0 as the data actually observed. A very low p-value is taken as evidence for discovery; the threshold is often taken to be $p = 2.85 \times 10^{-7}$, corresponding to a 5σ fluctuation in one direction of a Gaussian variable.

In Bayesian statistics, all of one's knowledge about a model is contained in the posterior probabilities. Thus one could reject the null hypothesis if its posterior probability $P(H_0|\mathbf{x})$ is sufficiently small. The difficulty here is that $P(H_0|\mathbf{x})$ is proportional to the prior probability $P(H_0)$, and there will not be a consensus about the prior probabilities for the non-existence of a given new phenomenon. Nevertheless one can construct a quantity called the Bayes factor (described below), which plays the role of the frequentist p-value and is independent of $P(H_0)$.

Consider a scenario where we have two models, H_0 and H_1, e.g., the 'Standard Model minus Higgs' and the 'Standard Model with Higgs'.[1] Models H_0 and H_1 are described by vectors of parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, respectively. Some of the components will be common to both models and others may be distinct. The full prior probability can be written

$$\pi(H_i, \boldsymbol{\theta}_i) = p_i \pi_i(\boldsymbol{\theta}_i) \ , \tag{7}$$

Here $p_i = P(H_i)$ is the overall prior probability for H_i, $i = 0, 1$, and $\pi_i(\boldsymbol{\theta}_i)$ is the normalized pdf of its parameters.

---

[1] The Standard Model without the Higgs mechanism may not be a well defined theory, but we can regard it as an effective model in which all events except those with real Higgs production are allowed.

We would claim discovery that H_1 is true if we find it has a high posterior probability $P(H_1|\mathbf{x})$. Instead of quoting the posterior probability itself, it is equivalent to give the posterior odds, i.e., the probability of H_1 divided by that of its complement H_0,

$$\Omega_{10} = \frac{P(H_1|\mathbf{x})}{P(H_0|\mathbf{x})} . \tag{8}$$

We can compare this with the prior odds,

$$\omega_{10} = \frac{P(H_1)}{P(H_0)} = \frac{1 - p_0}{p_0} . \tag{9}$$

The ratio of posterior to prior odds is the Bayes factor,

$$B_{10} = \frac{\Omega_{10}}{\omega_{10}} . \tag{10}$$

Equivalently one can define the Bayes factor using the odds of the null hypothesis H_0 relative to the alternative H_1, i.e., $B_{01} = 1/B_{10}$. The Bayes factor says how the odds of one hypothesis compared to its complement change in the light of the data. This quantity is independent of the prior probability p_0 (shown below) and thus provides a numerical measure of evidence supplied by the data in favour of one hypothesis over the other. One would use a high value of B_{10} as indicating discovery of the new phenomenon (the complement, H_1), in a manner analogous to how one would regard a low p-value in frequentist statistics [1]. The values in Table 1 been proposed [10] to give a rough idea to the importance that one might attach to different values of the Bayes factor.

Table 1: Interpretation of the Bayes factor B_{10} (from [10]).

B_{10}	Evidence against H_0
1 to 3	Not worth more than a mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

The Bayes factor B_{10} can be written

$$B_{10} = \frac{P(H_1|\mathbf{x})}{P(H_1)} \bigg/ \frac{P(H_0|\mathbf{x})}{P(H_0)} . \tag{11}$$

In general each hypothesis is characterized by a set of internal parameters $\boldsymbol{\theta}_i$, $i = 0, 1$, some components of which may be common to the two models. To find the probabilities needed in equation (11) we must integrate over these parameters, e.g.,

$$P(H_i|\mathbf{x}) = \int P(H_i, \boldsymbol{\theta}_i|\mathbf{x}) \, d\boldsymbol{\theta}_i , \tag{12}$$

Using Bayes theorem we can write this as

$$P(H_i|\mathbf{x}) = \frac{\int L(\mathbf{x}|H_i, \boldsymbol{\theta}_i) p_i \pi_i(\boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i}{P(\mathbf{x})} , \tag{13}$$

5

where $p_i = P(H_i)$ is the prior probability for the model H_i as a whole, and $\pi_i(\boldsymbol{\theta}_i)$ is the prior pdf for its parameters. The Bayes factor can therefore be written

$$B_{10} = \frac{\int L(\mathbf{x}|H_1, \boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}{\int L(\mathbf{x}|H_0, \boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)\, d\boldsymbol{\theta}_0} \ , \tag{14}$$

which does not depend on the prior probability p_0. This is an important property of the Bayes factor. One can imagine a relatively broad spectrum of opinions regarding the prior probability of some new phenomenon. But the Bayes factor only says how the odds ratio evolves in the light of the data, and this factor is the same regardless of one's subjective prior probabilities.

The Bayes factor does require prior probabilities for all internal parameters of a model, e.g., one needs the functions $\pi_i(\boldsymbol{\theta})$. In some Bayesian analyses it is acceptable to take an unnormalizable function for the prior (an improper prior) as long as the product of likelihood and prior can be normalized. Improper priors are only defined up to an arbitrary multiplicative constant. If some of the nuisance parameters λ are common to the two models H_0 and H_1, then we may use improper priors for these, since the arbitrary constant would cancel in the Bayes factor. But if a parameter appears in model H_1 but not in H_0 then the arbitrary constant in $\pi_1(\boldsymbol{\theta}_1)$ would not cancel in the Bayes factor, which would then be ill defined. So all parameters that are not common to both models must be described by normalized priors.

Both integrals in equation (14) are of the form

$$m = \int L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \ , \tag{15}$$

which is called the marginalized likelihood. Here $\boldsymbol{\theta}$ is taken to indicate all of the internal parameters of the hypothesis in question. Note that the normalizing factor in the denominator of Bayes' theorem (1) is also of this form, i.e., $m = P(\mathbf{x})$ gives the prior probability of the data $\mathbf{x}$.

1.5 Numerical determination of Bayes factors

Given that we know how to sample from the posterior pdf using MCMC, one might think it would be simple to obtain from this the marginal likelihoods and from these the Bayes factor. In fact estimating the integral (15) entails a number of new computational challenges.

It is difficult to give a single numerical method appropriate to computing Bayes factors in all problems. In the remainder of this section we summarize several methods.

1.5.1 Harmonic mean estimator

If we consider only one of the models, then we can rearrange Bayes' theorem as

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})} \ . \tag{16}$$

Since $\pi(\boldsymbol{\theta})$ is a pdf normalized to unity we can integrate both sides of (16) to find

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L] \ , \tag{17}$$

where $E_p[\cdot]$ denotes the expectation value with respect to the posterior pdf $p(\boldsymbol{\theta}|\mathbf{x})$. So can take the sequence of $\boldsymbol{\theta}$ values generated using MCMC as described in Section 1.2 and use these to estimate the mean of $1/L$, and then the inverse of this, i.e., the harmonic mean [11] of L, gives the estimate of m.

It turns out that the harmonic mean estimator for m is numerically unstable, since the posterior can occasionally generate $\boldsymbol{\theta}$ values in regions of very small L, and these cause large fluctuations in the average of $1/L$. Formally the variance of the average for any finite number of MCMC steps is infinite [10]. Nevertheless, this method can be used to obtain a rough determination of the Bayes factor, and it is often sufficiently accurate (say, within a factor of two) to be useful as a first estimate.

1.5.2 Stabilized harmonic mean estimator

One can improve upon the harmonic mean estimate if one has a normalized pdf $f(\boldsymbol{\theta})$ which can be evaluated numerically at an arbitrary point $\boldsymbol{\theta}$ [10, 12]. Rearranging Bayes' theorem and multiplying both sides by $f(\boldsymbol{\theta})$ we find

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m} \ . \tag{18}$$

Integrating both sides over $\boldsymbol{\theta}$ gives

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}p(\boldsymbol{\theta}|\mathbf{x}) = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right] \ . \tag{19}$$

Therefore we can generate $\boldsymbol{\theta}$ values from the posterior using MCMC and compute the average of $f/L\pi$. To improve the numerical stability of the estimator we must choose the pdf $f(\boldsymbol{\theta})$ to have tails that fall off faster than those of the posterior density so as to avoid large values of $f/L\pi$. A truncated multivariate Gaussian centred about the posterior has been suggested [13]. Although this method can improve somewhat the convergence properties of the estimated marginal likelihood, for high-dimensional problems it can be difficult to find an appropriate density $f(\boldsymbol{\theta})$ [10]. Note that the harmonic mean is simply the special case of $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

1.5.3 Importance sampling

Suppose we have a normalized pdf $f(\boldsymbol{\theta})$ which we can evaluate numerically and from which we can also generate a Monte Carlo sample, such as a multivariate Gaussian. We can write the marginalized likelihood (15) as

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}f(\boldsymbol{\theta})\,d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right] \ , \tag{20}$$

where $E_f[\cdot]$ denotes the expectation value with respect to the density $f(\boldsymbol{\theta})$. In importance sampling, one computes the expectation value of $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ by sampling $\boldsymbol{\theta}$ from a suitable density $f(\boldsymbol{\theta})$ and computing the average of $L\pi/f$.

The convergence is fastest if the function $f(\boldsymbol{\theta})$ approximates the shape of $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. One can use, for example, a multivariate Gaussian whose mean and covariance are estimated from the posterior pdf $p(\boldsymbol{\theta}|\mathbf{x}) \propto L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ using, .e.g, MINUIT.

1.5.4 Parallel tempering

Another method for computing the marginal likelihood exploits a procedure known as parallel tempering [4]. One begins by defining

$$Z(\beta) = \int L^{\beta}(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \ , \tag{21}$$

which is called the partition function in analogy with statistical mechanics. At $\beta = 0$ and $\beta = 1$ we have

$$Z(0) = \int \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = 1 \ , \tag{22}$$

$$Z(1) = \int L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = m \ . \tag{23}$$

The strategy is to find $m = Z(1)$ by expressing its logarithm as

$$\ln Z(1) = \ln Z(0) + \int_0^1 \frac{d \ln Z}{d\beta} \, d\beta \ , \tag{24}$$

and therefore

$$m = \exp\left[\int_0^1 \frac{1}{Z}\frac{dZ}{d\beta} \, d\beta\right] \ . \tag{25}$$

By first writing $Z(\beta)$ as

$$Z(\beta) = \int \exp\left[\ln \pi(\boldsymbol{\theta}) + \beta \ln L(\mathbf{x}|\boldsymbol{\theta})\right] \, d\boldsymbol{\theta} \tag{26}$$

we find the derivative needed in equation (25) to be

$$\begin{aligned}
\frac{dZ}{d\beta} &= \int \ln L(\mathbf{x}|\boldsymbol{\theta}) \, \exp\left[\ln \pi(\boldsymbol{\theta}) + \beta \ln L(\mathbf{x}|\boldsymbol{\theta})\right] \, d\boldsymbol{\theta} \\
&= \int \ln L(\mathbf{x}|\boldsymbol{\theta}) \, L^{\beta}(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \ .
\end{aligned} \tag{27}$$

The full integrand needed for equation (25) is therefore

$$\frac{1}{Z}\frac{dZ}{d\beta} = \frac{\int \ln L(\mathbf{x}|\boldsymbol{\theta}) \, L^{\beta}(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int L^{\beta}(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} = E_{\beta}\left[\ln L(\mathbf{x}|\boldsymbol{\theta})\right] \ , \tag{28}$$

where $E_{\beta}[\cdot]$ denotes the expectation value with respect to the modified density $p_{\beta} \propto L^{\beta}\pi$, rather than the posterior density, which corresponds to $\beta = 1$.

The same MCMC technology that allows us to simulate from the posterior density can also be used to generate a sample following the modified density $p_{\beta} \propto L^{\beta}\pi$. Thus for any given β we can generate an MCMC sample and estimate (28) using the average value of $\ln L(\mathbf{x}|\boldsymbol{\theta})$. This can be done for, say, 10 to 20 values of β and then the integral (25) can be found using a spline fit.

8

2 Bayesian methods for Higgs search

In this section we will apply the formalism described above to the specific case of a search for the Standard Model Higgs boson. In Section 2.1 the case of a single decay channel based on event counting is considered. In Section 2.2 this is extended to a combined analysis of multiple channels.

There is some freedom as to how one models the set of measurements. One way to would be to regard the Higgs mass m_{H} as the parameter of interest, and $m_{\mathrm{H}} = \infty$ is effectively a label for the null hypothesis. One would then assume the Standard Model relation between the Higgs cross section and its mass, and make inferences about the mass.

Alternatively we can choose to not impose the SM relation between cross section and mass, but rather to choose a fixed mass and regard the cross section as the parameter of interest. Once we have found the posterior probability for the cross section we repeat the entire analysis for all masses, moving in small steps over the mass region of interest.

For any given mass one can construct, say, the 95% upper limit on the cross section by integrating the posterior density (cf. equation (6)). At low masses these limits will be below the predicted cross section, which is to say that these mass values are excluded. The lower limit on the mass is obtained from the point where the prediction is equal to the upper limit on the cross section. The result packaged in this way can be used to place constraints the Standard Model or on Higgs-like particle production in models with non-SM couplings.

If we find a small Bayes factor for any mass then this provides evidence for a discovery. One can then impose the SM relation between cross section and mass to estimate the Higgs mass.

2.1 Single channel with event counting

Consider the case where one observes n events in a given region where one is looking for evidence of the signal process (Higgs production for a given decay channel). The expectation value of n is

$$E[n] = s + b , \tag{29}$$

where s is the expected number of signal events and b is the expected number of background events. In general we can express s as

$$s = \sigma_{\mathrm{s}} \mathcal{B} \varepsilon_{\mathrm{s}} L , \tag{30}$$

where σ_{s} is the signal cross section, $\mathcal{B}$ is the branching ratio for the decay channel in question, ε_{s} is the signal efficiency and L is the integrated luminosity.

The number of events n is assumed to follow a Poisson distribution, i.e., its probability is

$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} . \tag{31}$$

Here only s and b have been listed as parameters, but we would usually regard the cross section σ as the parameter of interest and ε, $\mathcal{B}$, L and b would be nuisance parameters.

Suppose that to determine the expected number of background events b one carries out a subsidiary measurement using either a sideband or some control region where no signal is

expected. Suppose this yields a single Poisson distributed value m with mean value τb. Here τ is a parameter that effectively scales the size of the sideband region to that of the signal. For now let us assume that it can be determined, e.g., using Monte Carlo, with negligible uncertainty.

Note that if we take the (improper) prior for b to be constant for $b \geq 0$ and zero otherwise, then the posterior for b after carrying out the subsidiary measurement of m only is found from Bayes' theorem to be

$$P(b|m,\tau) \propto P(m|b,\tau) = \frac{(\tau b)^m}{m!}e^{-(\tau b)}, \quad b \geq 0, \tag{32}$$

Integrating over b to normalize (32) to unity gives a Gamma distribution,

$$P(b|m,\tau) = \frac{\tau}{m!}(\tau b)^m e^{-\tau b}, \quad b \geq 0, \tag{33}$$

which has a mode at m/τ, a mean of $E[b] = (m+1)/\tau$ and a standard deviation $\sigma_b = \sqrt{m+1}/\tau$. Under some circumstances one might estimate the background not from a sideband measurement but rather using some other indirect information. One might model this using a Gamma prior with a mean and standard deviation equal to the background estimate and its 1σ error, respectively. For now, however, let us only consider the Poisson measurement m that provides our information about b.

Since n and m are independent we can write the likelihood function as

$$L(n,m|s,b) = \frac{(s+b)^n}{n!}e^{-(s+b)}\frac{(\tau b)^m}{m!}e^{-(\tau b)}. \tag{34}$$

Using this with a uniform prior $\pi_b(b)$ for $b \geq 0$ and zero otherwise, the Bayes factor B_{10} becomes

$$B_{10} = \frac{\int\int L(n,m|s,b)\pi_b(b)\pi_s(s)\,db\,ds}{\int L(n,m|s=0,b)\pi_b(b)\,db} \tag{35}$$

$$= \frac{\int\int (s+b)^n(\tau b)^m e^{-s-(1+\tau)b}\pi_s(s)\,db\,ds}{\int b^n(\tau b)^m e^{-(1+\tau)b}\,db}. \tag{36}$$

For the case of $s = 0$, we can compute the marginalized likelihood in closed form:

$$\int_0^\infty \frac{b^n e^{-b}}{n!}\frac{(\tau b)^m e^{-\tau b}}{m!}\,db = \frac{\tau^m(n+m)!}{n!m!(1+\tau)^{n+m+1}}. \tag{37}$$

This provides a useful point of reference for checking the numerical calculations.

2.2 Combination of multiple channels with event counting

Now consider N independent search channels. In this section will compute a single Bayes factor that compares the null (no Higgs) hypothesis H_0 with an alternative H_1 that assumes the existence of a Higgs boson at a given mass m_H with Standard Model branching ratios into all of the channels. We can consider the hypothesis where the Higgs production rate is equal to the SM prediction for the chosen m_H, but we will also introduce a global parameter that

will allow us to consider different Higgs production rates (but always assuming SM branching ratios).

Suppose for the ith channel we define a signal region in which we find n_i events, and we regard each n_i as an independent Poisson variable with mean $s_i + b_i$, where s_i and b_i are the contributions from signal and background. The collection of measurements $\mathbf{n} = (n_1, \ldots, n_N)$ has the joint probability

$$P(\mathbf{n}|\mathbf{s}, \mathbf{b}) = \prod_{i=1}^{N} \frac{(s_i + b_i)^{n_i}}{n_i!} \, e^{-(s_i+b_i)} \, . \tag{38}$$

Let us assume that the expected number of signal events can be written as

$$s_i = \mu \sigma_{\mathrm{SM}} \mathcal{B}_i \varepsilon_{s,i} L_i \equiv \mu \varphi_i \, , \tag{39}$$

where σ_{SM} and $\mathcal{B}_i$ are the Standard Model values for the cross section and branching ratio, $\varepsilon_{s,i}$ is the signal efficiency and L_i is the integrated luminosity, which could be different for each channel. Here μ is a global signal-strength parameter, defined such that $\mu = 1$ gives the Standard Model prediction for s_i. Further we have defined $\varphi_i = \sigma_{\mathrm{SM}} \mathcal{B}_i \varepsilon_{s,i} L_i$, so that systematic uncertainties in any of the factors can be treated as an uncertainty in φ_i.

As in the single-channel case suppose we have a set of independent subsidiary measurements $\mathbf{m} = (m_1, \ldots, m_N)$ to determine the background values. Their joint probability is

$$P(\mathbf{m}|\mathbf{b}, \boldsymbol{\tau}) = \prod_{i=1}^{N} \frac{(\tau_i b_i)^{m_i}}{m_i!} \, e^{-\tau_i b_i} \, , \tag{40}$$

where as before we will assume that the τ_i can be determined with negligible uncertainty.

The factors that appear in φ_i (SM cross section, branching ratio, efficiency and luminosity) will in general have some uncertainty. Let us suppose that we can quantify this by specifying a prior probability density $\pi_{\varphi,i}(\varphi_i)$. The details of this function could in general be complicated and depend on many factors. As a first approximation one might know that it is centred about the nominal value and has a standard deviation $\sigma_{\varphi,i}$. One could model such a case as using a Gamma distribution with a mean equal to the nominal value of φ_i and with a standard deviation equal to the estimated standard error $\sigma_{\varphi,i}$.

A Gamma distribution for the φ factor of an individual channel is

$$\pi_\varphi(\varphi) = \frac{a(a\phi)^{b-1} e^{-a\varphi}}{\Gamma(b)} \, , \tag{41}$$

where the parameters a and b are related to the mean $E[\varphi]$ and variance $V[\varphi]$ by $E[\varphi] = b/a$ and $V[\varphi] = b/a^2$. Most often one would be provided with a nominal value φ_0 and relative error $r_\varphi = \sigma_\varphi/\varphi_0$. Taking these as the mean and standard deviation of the prior gives for the parameters of the Gamma distribution,

$$a = \frac{1}{\varphi_0 r_\varphi^2} \, , \tag{42}$$

$$b = \frac{1}{r_\varphi^2} \, . \tag{43}$$

11

For the joint pdf of the set of factors $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_N)$ one may wish to consider correlations between some of the terms. This would be the case, for example, for the luminosity factors. For the present example, however, let us assume that the uncertainties in the φ values are all independent so that we have a joint pdf

$$\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) = \prod_{i=1}^{N} \frac{a_i(a_i\varphi_i)^{b_i-1}e^{-a_i\varphi_i}}{\Gamma(b_i)} \; , \tag{44}$$

where the a_i and b_i are related to the corresponding nominal values and relative errors as described above.

Now suppose we carry out the measurement and find data $\mathbf{n} = (n_1, \ldots, n_N)$ and $\mathbf{m} = (m_1, \ldots, m_N)$. Using the ingredients above in equation (14) gives the Bayes factor B_{01},

$$B_{10} = \frac{\int \int \int L(\mathbf{n}, \mathbf{m}|\mu, \mathbf{b}, \boldsymbol{\varphi}) \, \pi_\mu(\mu)\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\pi_{\mathbf{b}}(\mathbf{b}) \, d\mu \, d\boldsymbol{\varphi} \, d\mathbf{b}}{\int \int L(\mathbf{n}, \mathbf{m}|\mu = 0, \mathbf{b}, \boldsymbol{\varphi}) \, \pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\pi_{\mathbf{b}}(\mathbf{b}) \, d\boldsymbol{\varphi} \, d\mathbf{b}} \; . \tag{45}$$

2.3 Numerical example with multiple channels

As a numerical test, values for expected numbers of signal and background events for a Higgs with a mass of 130 GeV were taken from [15]. The relevant numbers are shown in Table 2, where the values for the channel $H \to WW^* \to l\nu\,jj + X$ have been decreased by a factor of 3 relative to the publication so that all channels correspond to an integrated luminosity of 10 fb^{-1}.

Table 2: Expected numbers of signal and background events, s and b, for three decay channels with $m_{\mathrm{H}} = 130$ GeV, all scaled to an integrated luminosity of 10 fb^{-1} (from [15] Table 7).

Channel	s	b
$H \to WW^* \to e\mu + X$	12.3	9.2
$H \to WW^* \to ee/\mu\mu + X$	11.7	10.1
$H \to WW^* \to l\nu\,jj + X$	1.5	2.0

For this example a desired relative uncertainty in the background b was chosen, and this was translated into a corresponding size of the sideband through the parameter τ. Similarly, a value of the relative uncertainty in φ was specified and this was translated into parameter values of a Gamma distribution using equations (42) and (43).

In this example, a single data set was created by taking integer values n_i closest to the $s_i + b_i$ from Table 2. In the real experiment of course one uses whatever n_i Nature delivers. Alternatively one could generate Poisson distributed data sets based on specific values of s_i and b_i to investigate the expected sensitivity.

Importance sampling was used to compute the Bayes factor comparing the no Higgs hypothesis H_0 to the alternative H_1 of a specific value of the strength parameter μ. For this example, all of the methods for determining the Bayes factor other than importance sampling indicated some numerical difficulties (still under investigation). The results are shown in Fig. 1 for $\sigma_\varphi/\varphi = 0.1$ for several values of the background uncertainty σ_b/b. Figure 2 shows the corresponding quantity for $\sigma_b/b = 0.1$ with several values of σ_φ/φ.

Since the data values n_i and m_i were chosen to correspond to s_i and b_i (rounded to the nearest integer), one finds as expected the highest value of B_{10} for $\mu \approx 1$. Although a 30%
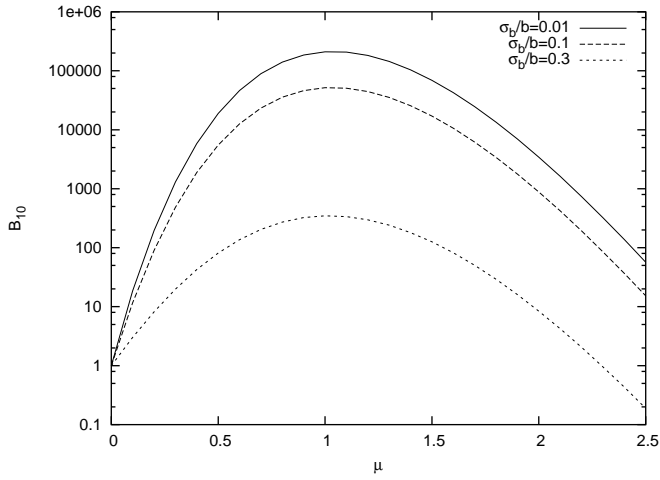
Figure 1: The Bayes factor B_{10} versus the global strength parameter μ with $\sigma_\varphi/\varphi = 0.1$ for several values of the background uncertainty σ_b/b.
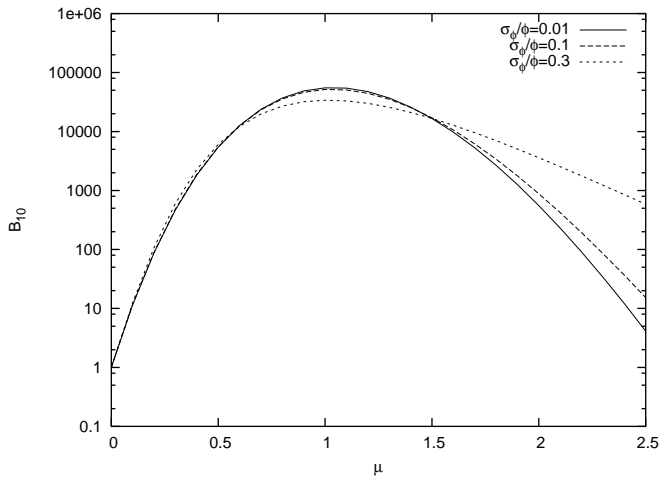

Figure 2: The Bayes factor B_{10} versus the global strength parameter μ with $\sigma_b/b = 0.1$ for several values of σ_φ/φ.

13

uncertainty in the signal efficiency has only a minor effect on the peak value of B_{10}, a 30% uncertainty in the background reduces the peak B_{10} from roughly 10^5 to 10^2.

In the example above, the data set was created to correspond to the presence of a signal by taking n_i equal to the integer closest to $\mu\varphi_i + b_i$. We can also consider data sets that one would expect in the absence of signal, i.e., n_i close to b_i. Such a data set was created by taking the integers closest to the background values from Table 2: $n_1 = 9$, $n_2 = 10$, $n_3 = 2$. In this case we find a Bayes factor that drops rapidly for increasing μ, with $B_{10} = 6.4 \times 10^{-5}$ at $\mu = 1$, as see in Fig. 3.
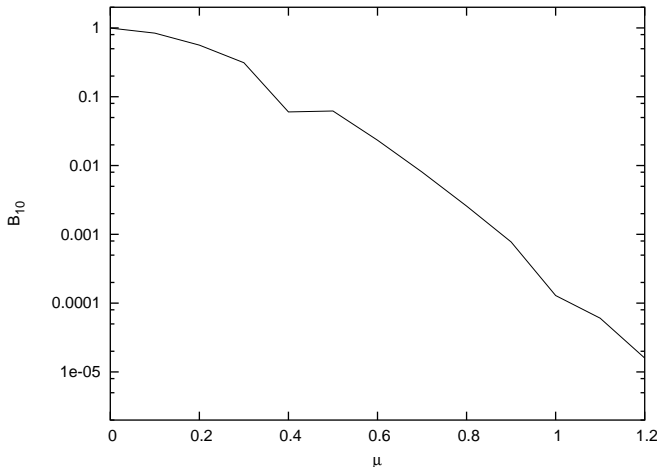

Figure 3: The Bayes factor B_{10} versus the global strength parameter μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$, $\sigma_b/b = 0.1$, $\sigma_\varphi/\varphi = 0.1$).

The posterior distribution of μ for this data set is shown in Fig. 4, based on an MCMC sample with 10^6 steps and an improper uniform prior $\pi_\mu(\mu) = 1$, $\mu \geq 0$. The curves correspond to three different values of the background uncertainty, σ_b/b; all have $\sigma_\varphi/\varphi = 0.1$. The 95% CL upper limits come out to $\mu_{\rm up} = 0.418$, 0.429 and 0.493 for $\sigma_b/b = 0.01$, 0.1 and 0.3, respectively. Figure 5 shows the corresponding curves for $\sigma_b/b = 0.1$ and several values of σ_φ/φ. The 95% CL upper limits come out to $\mu_{\rm up} = 0.424$, 0.429 and 0.474 for $\sigma_\varphi/\varphi = 0.01$, 0.1 and 0.3, respectively.

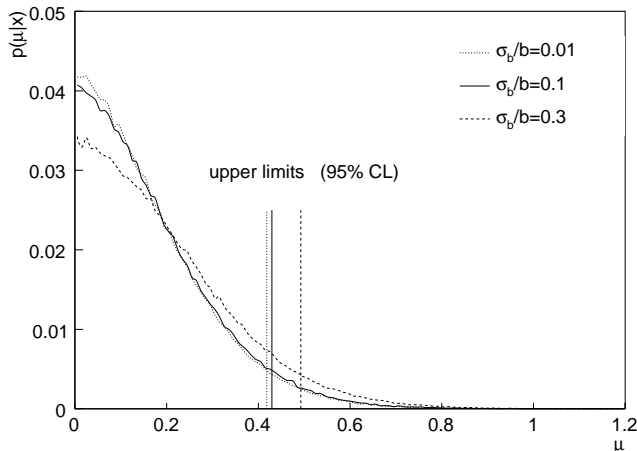

Figure 4: The posterior distribution of μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$) for $\sigma_\varphi/\varphi = 0.1$ and several values of σ_b/b.

Note that the improper prior used for the limit cannot be used to find a Bayes factor. There the prior is only defined up to an arbitrary constant, and this would not cancel in B_{10}.
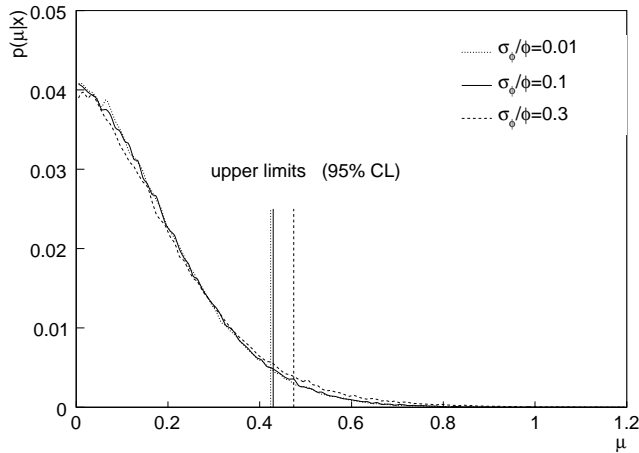
Figure 5: The posterior distribution of μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$) for $\sigma_b/b = 0.1$ and several values of σ_φ/φ.

One could use, e.g., a normalized exponential prior

$$\pi_\mu(\mu) = \frac{1}{\mu_0} e^{-\mu/\mu_0}\,, \quad \mu \geq 0\,, \tag{46}$$

and choose μ_0 to be some value that is large compared to the Standard Model expectation of $\mu = 1$. While this is again acceptable for the posterior pdf and therefore also the limit, the Bayes factor would simply go to zero in the limit of larger μ_0, since the fraction of area under $\pi_\mu(\mu)$ that is compatible with the data decreases as μ_0 increases; the bulk of the prior probability becomes concentrated at unrealistically high values of μ.

In some sense the 'look-elsewhere effect' is taken into account by focusing on the alternative H_1 for which $\mu = 1$, rather than on that value of μ which maximizes the Bayes factor.

References

[1] Glen Cowan, *Statistical methods for the ATLAS Higgs search*, contribution to the ATLAS Statistics forum, 11 April, 2007.

[2] J. Ghosh, M. Delampady and T. Samanta, *An Introduction to Bayesian Analysis: Theory and Methods*, Springer, 2006.

[3] Anthony O'Hagan, *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*, Wiley, 1994.

[4] Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

[5] Joel Heinrich, *Bayesian limit software: multi-channel with correlated backgrounds and efficiencies*, CDF/MEMO/STATISTICS/PUBLIC/7587 (2005).

[6] Joel Heinrich et al., *Interval estimation in the presence of nuisance parameters. 1. Bayesian approach*, CDF/MEMO/STATISTICS/PUBLIC/7117, physics/0409129 (2004).

[7] Luc Demortier, *A Fully Bayesian Computation of Upper Limits for Poisson Processes*, CDF/MEMO/STATISTICS/PUBLIC/5928 (2004).

[8] Christian P. Robert and George Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, 2004.

[9] Jun S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2001.

[10] Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

[11] M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

[12] A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

[13] J. Geweke, *Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication*, Econometric Reviews 18 (1999) 1-126.

[14] Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

[15] S. Asai et al., *Prospects for the Search for a Standard Model Higgs Boson in ATLAS using Vector Boson Fusion*, Eur. Phys. J. C32S2 (2004) 19-54; hep-ph/0402254.