A Bayesian Approach for Hypothesis Testing in Particle Physics

Shirin Golchi*1 and Richard Lockhart2

1Applied Statistics Centre, Columbia University, New York, NY, USA 10027
2Department of Statistics and Actuarial Science, Simon Fraser university, Burnaby, BC, Canada V5A1S6

Abstract

The statistical procedure used in a family of problems in particle physics, similar to the search for the Higgs boson, is investigated in this paper. We propose a Bayesian hierarchical model to make inference while addressing treatment of nuisance parameters and introduce a Bayesian decision making procedure that combines the two steps of the current method (discovery and exclusion) into one, while enabling us to control global frequency theory error rates and use the information provided by the theory.

Keywords: Bayes rule, Higgs boson, hierarchical model, linear loss function

1 Introduction

The Standard Model (SM) of particle physics is a theory that describes the dynamics of subatomic particles. The Higgs particle is an essential component of the SM; its existence explains why other elementary particles are massive. The existence of the Higgs boson needs to be confirmed by experiment. The Large Hadron Collider (LHC) at the European Organization for Nuclear Research, known as CERN, is a high energy collider specifically designed and constructed to detect the Higgs particle. Two beams of protons circulating at very high speeds in the LHC collide inside two detectors (Atlas and CMS). Collisions between a proton in one beam and a proton in the other beam result in generation of new particles, possibly including the Higgs boson; each such collision is an event. Some of the particles generated can be tracked and measured in the detectors. However, the Higgs particle, if generated, decays extremely quickly into other known SM particles and cannot, therefore,

*e-mail: sg3252@columbia.edu
be detected directly. Instead, the existence of the Higgs particle must be inferred by looking for those combinations of detectable particles that are predicted by the SM.

Once a Higgs particle has been created (by one of several ‘production mechanisms’) in a proton-proton collision, there are several different processes, called ‘decay modes’, through which the particle may decay. The decay process can be reconstructed based on the detected collision byproducts. Events with reconstructed processes that match one of the possible Higgs decay modes and pass other selection criteria (called cuts) are recorded as “Higgs events” and the invariant mass of the unobserved particle is computed from the reconstruction. A histogram of the invariant mass is then created for each decay mode. However, there are other processes, not involving the Higgs boson, that can result in the generation of Higgs event byproducts which also pass the cuts; these are called background events. Thus the histogram created is either a mixture of background events and events in which a Higgs particle was created or just a histogram of background events if the Higgs particle does not exist.

Luckily, the SM predicts, as a function of the mass of the Higgs particle and the energy of the colliding beams, the expected rate at which events generate Higgs particles (a quantity proportional to the so-called Higgs cross section). It also predicts, as a function of the same parameters, the probability that the particle will decay by a given decay mode and produce by-products which pass the cuts. Effectively then, the SM predicts that if the Higgs particle exists and has mass \( m \) then there will be a bump on the histogram of invariant masses whose size and shape are completely predicted by unknown mass, \( m \), and other measured quantities such as beam energy. The statistical problem is to determine whether or not such a bump exists and at what mass the bump is centred if it does exist.

Particle searches generally have two stages. The first stage is discovery, or search while the second is called exclusion. For each mass in a range of possible masses the null hypothesis that there is no Higgs particle is tested against the alternative that it exists and has this specific mass. A p-value, called a local p-value, is computed for each mass; if the smallest p-value is less than \( 1 - \Phi(5) \) (a 5\( \sigma \) effect—\( \Phi \) is the standard normal cumulative density function) then the particle is declared to have been ‘observed’. If not then the second stage, exclusion, is pursued. For each mass in the range under consideration the null hypothesis that the particle exists, at this mass, with the predicted cross section, is tested against a background only null hypothesis.

While the main goal is to discover the Higgs particle and the parameter of interest is the mass of the Higgs boson, denoted by \( m_H \), the local hypothesis tests (which are likelihood ratio tests) focus on the cross section at each mass. The background model parameters and any other unknown parameters of the signal function are treated as nuisance parameters. We treat the cross section at mass value, \( m \), as a parameter denoted by \( c_m \) and denote the cross section predicted by the SM by \( c_m^{SM} \).

We propose a methodology to combine discovery and exclusion. We test the global null hypothesis that there is no Higgs particle with a mass in some range over which a search is carried out and rule out as many masses as we can as possibilities for the mass of the Higgs particle.
In our methodology, we use the specific predictions provided by the standard model. This contrasts with current practice. In the searches for the Higgs particle described in CMS Collaboration (2012) and ATLAS Collaboration (2012), in the discovery phase, the cross section was replaced with the value predicted by SM multiplied by an unknown unitless scaling factor, generally denoted by \( \mu \). The standard model then predicts that \( \mu = 1 \); the null hypothesis that there is no Higgs is represented by \( \mu = 0 \) but the procedures described in CMS Collaboration (2012) and ATLAS Collaboration (2012) treat the alternative to \( \mu = 0 \) as \( \mu > 0 \) (since this is equivalent to \( H_A : c_m > 0 \) we will not use the notation \( \mu \) in the sequel). The justification for ignoring in this way information provided by the theory is to investigate scenarios where there is evidence for existence of undiscovered phenomena different from the one suggested by theory and to guard against a situation where a bump of unpredicted size is found on the histogram due to misunderstandings about the background process.

In order to fix some notation we now give some further details of the two stages for current practice. In the process we want to highlight some key weaknesses of the existing method.

**Discovery:** For each mass value, \( m \in (x_0, x_n) \), in the search window a local likelihood ratio test (see Cowan et al. (2011) for details) is performed at the significance level, \( \alpha_0 = 1 - \Phi(5) = 3 \times 10^{-7} \),

\[
H_0 : c_m = 0 \quad \text{vs.} \quad H_A : c_m > 0.
\]

Note that the cross section predicted by the SM, \( c^{SM}_m \), is not used under the alternative. Discovery is announced if the p-value associated with any of the local tests (the minimum local p-value), is below \( \alpha_0 \) and the mass of the Higgs particle is estimated as the corresponding mass value. This testing procedure is equivalent to rejecting the null hypothesis if any of a family of local test statistics, \( q(m) \), (e.g., log likelihood ratios) indexed by the unknown mass \( m \) is larger than a predefined level \( \kappa \). Gross and Vitells (2010) proposed a method for estimating the “global p-value” of this testing procedure as the null probability that the maximum of the local test statistics is greater than the observed maximum. Their method is based on the results of Davies (1987) and gives:

\[
P_G = \Pr_{H_0}(\max_m q(m) > \kappa) \\
\approx \Pr_{H_0}(\chi^2_d > \kappa) + \mathbb{E}(N(\kappa))
\]

where \( \kappa \) in this case is the observed maximum statistic, \( \chi^2_d \), denoting a chi-squared distribution with \( d \) degrees of freedom, is the null distribution of \( q(m) \) and \( \mathbb{E}(N(\kappa)) \) is the expected number of upcrossings of the level \( \kappa \) by the process \( q(m) \). Using the method of Gross and Vitells (2010) one can estimate the (global) type I error rate associated with the discovery procedure that is based on the local tests,

\[
\alpha_G = \Pr_{H_0}(\max_m q(m) > \kappa_{\alpha_0}) \\
\approx \Pr_{H_0}(\chi^2_d > \kappa_{\alpha_0}) + \mathbb{E}(N(\kappa_{\alpha_0})) \\
= \alpha_0 + \mathbb{E}(N(\kappa_{\alpha_0}))
\]
where $\kappa_{\alpha_0}$ is the $\chi^2$ quantile corresponding to $\alpha_0$. Since $E(N(\kappa)) \geq 0$, the actual global type I error rate is larger than the controlled local type I error rates. The size of the difference depends on the specific statistical model and cannot be controlled in advance.

**Exclusion:** If enough evidence for discovery is not provided in the first stage of the analysis, further investigation is carried out to exclude regions of $m$ as unlikely values for the mass of the Higgs boson. The theoretical cross section is tested at $\alpha_2 = 0.05$ significance level at the exclusion step (Cowan et al., 2011),

$$H_0 : c_m = c_{m}^{SM} \quad \text{vs.} \quad H_A : c_m < c_{m}^{SM}.$$  

Any mass whose corresponding theoretical cross section is rejected is excluded from the range of possible masses for the Higgs particle.

A number of questions/criticisms arise regarding the existing procedure:

- What is the cost of ignoring the theoretical cross section at the discovery stage in terms of loss in sensitivity?
- What is the global type I error rate associated with the local discovery procedure?
- What would the decision be if more than one local p-value fell below the local significance level at the discovery step?
- How is the exclusion significance level justified?
- Why is a smaller cross section than the one predicted by theory eligible at the discovery step when it might well be excluded when testing for exclusion? (Of course current practice prevents this contradiction from arising by the expedient of not doing exclusion once discovery has been declared.)

In the following we propose a Bayesian procedure that,

- combines the discovery and exclusion into a single decision making step;
- enables the calculation and control of various frequency theory global error rates and can therefore be calibrated to have predetermined error rates;
- and takes advantage of the information provided by theory.

The rest of the paper is organized as follows; in Section 2, we introduce a Bayesian hierarchical model and describe inference for model parameters. Section 3 is dedicated to the Bayesian decision making procedure that while combining the discovery and exclusion steps can be calibrated to obtain desired frequency theory error rates. In Section 4, the proposed inference and decision making methods are applied to simulated Higgs data provided to us by the CMS Higgs group. Section 5 follows with concluding remarks.
2 Model and inference

In this section we introduce a Bayesian hierarchical model that captures the features of the Higgs problem. A sequential Monte Carlo (SMC) algorithm is used to make inference about the model parameters.

Suppose that the data, i.e., the invariant masses recorded by the detector, are a realization of a Poisson process whose intensity function is given by the sum of a background process \( \Lambda(x) \) and a signal function \( s(x; m) \). The shape of the signal function is known and its location is determined by the unknown parameter, \( m \in \mathcal{M} \), where \( \mathcal{M} = \{0\} \cup (x_0, x_n) \). The parameter, \( m \), is the unknown mass of the Higgs particle where \( m \in (x_0, x_n) \) means that the Higgs boson has a mass in the search window, \((x_0, x_n)\), while \( m = 0 \) refers to the case that the particle does not exist, at least not with a mass in \((x_0, x_n)\). We model the uncertainty about the background \( \Lambda(x) \) as a log-Gaussian process,

\[
\log \Lambda(x) \sim \mathcal{GP}(\mu(x), \rho(x, x')) \quad x \in (x_0, x_n).
\]

with known mean function, \( \mu(x) \), and covariance function given by,

\[
\rho(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{l^2}\right),
\]

where \( \sigma^2 \) is the variance parameter and \( l \) is the length scale that controls the smoothness of the background function.

We choose the signal function as a Gaussian probability density function with the location parameter \( m \) (in CMS Collaboration (2012) and ATLAS Collaboration (2012) a slightly more complex signal shape called the “crystal ball function” is used). Therefore, the signal function is given by,

\[
s_m(x) = 0 \mathbb{1}_{\{0\}}(m) + c_m \phi\left(\frac{x - m}{\epsilon}\right) \mathbb{1}_{(x_0, x_n)}(m),
\]

where \( c_m \) is a scaling constant chosen based on the known signal strength, \( \phi \) is the normal probability density function with known standard deviation, \( \epsilon \), that controls the spread of the signal function, and \( \mathbb{1}_A(\cdot) \) is the indicator function of set \( A \). Note that coefficient, \( c_m \), plays the same role as the cross section.

The use of finely binned data is common in the physics literature since the size of the data collected is often large. The likelihood of the binned data is given by,

\[
P(y|\Lambda, m) = \prod_{i=1}^{n} \frac{\exp(-\Gamma_i) \Gamma_i^m}{y_i!},
\]

where

\[
\Gamma_i = \int_{x_{i-1}}^{x_i} [\Lambda(x) + s_m(x)] dx.
\]

The grid \( x = (x_0, x_1, \ldots, x_n) \) is the vector of bin boundaries over the search window.
The posterior distribution of $\Lambda$ and $m$ given the data $y$ can be written as,

$$
\pi(m, \Lambda | y) = \frac{\pi(m, \Lambda) P(y | \Lambda, m)}{\int \int \pi(m, \Lambda) P(y | \Lambda, m) d\Lambda dm},
$$

(4)

where $\pi(m, \Lambda) = \pi(m) \pi(\Lambda)$ is the independent prior. The prior distribution, $\pi(m)$, is chosen from the family of "spike-and-slab" priors, i.e., a mixture of a point mass at $m = 0$ and a continuous distribution on $(x_0, x_n)$. With a multimodal prior distribution, sampling from (4) using regular Markov chain Monte Carlo is challenging since moving between the modes relying on random walk proposal strategies results in inefficiency or sometimes even infeasibility of sampling.

To overcome sampling problems we use an SMC algorithm (Del Moral et al., 2006). The SMC samplers are a family of algorithms that take advantage of a sequence of distributions that bridge between a distribution that is straightforward to sample from (for example the prior) and the target distribution. Particles are filtered through the defined sequence using importance sampling and re-sampling steps to eventually obtain a sample from the target distribution. In a common version of SMC the sequence of filtering distributions is defined by tempering the likelihood, i.e., the role of likelihood is induced in the model in a sequential manner. This is especially useful when introduction of the likelihood makes the posterior surface different from the prior in number and size of modes.

Let the filtering sequence of distributions be denoted by,

$$
\pi_0, \pi_1, \ldots, \pi_T.
$$

Using a temperature schedule $\{\tau_t, t = 0, \ldots, T\}$, the $t$th distribution in the sequence is defined as a power posterior,

$$
\pi_t = \pi(m, \Lambda)[P(y | \Lambda, m)]^{\tau_t},
$$

where

$$
0 = \tau_0 < \tau_1 < \ldots < \tau_T = 1.
$$

The SMC sampler comprises iterative steps of weighting and sampling. While the particles are moved towards the target distribution through re-sampling with weights calculated according to the current temperature, they are also moved towards higher probability regions under each distribution in the sequence, through a sampling step, to prevent particle degeneracy. These steps are explained in Algorithm 1.

In Algorithm 1 the simplified form of the incremental weights $\tilde{w}_t^i$ is the result of the choice of a MCMC kernel, $K_i$, in the sampling step. For more details about the SMC sampler and different choices of forward and backward kernels for the sampling step see Del Moral et al. (2006). In Section 4, we explain our specific choices of the inputs of the algorithm where we apply the above Bayesian hierarchical model to the simulated Higgs data set.

3 Bayesian decision making

In this section we consider the problem from a decision theoretic point of view. We define a linear loss function and derive the Bayes rule that can be used as an alternative to the
Algorithm 1 Sequential Monte Carlo

**Input:** a temperature schedule \( \{\tau_t, t = 0, \ldots, T\} \)
- a MCMC transition kernel \( K_t \)

1: Generate an initial sample \((m, \Lambda)_0^{1:N} \sim \pi_0\);
2: \( W_1^{1:N} \leftarrow \frac{1}{N} \);
3: for \( t := 1, \ldots, T - 1 \) do
   - \( W_i^t \leftarrow W_i^t \frac{\tilde{w}_i^t}{\sum \tilde{w}_i^t} \) where \( \tilde{w}_i^t = P(y|m_i^{(i)}, \Lambda_i^{(i)})^{\tau_t - \tau_{t-1}}, i = 1, \ldots, N; \)
   - Re-sample the particles \((m, \Lambda)_t^{1:N}\) with importance weights \( W_1^{1:N} \);
   - \( W_1^{1:N} \leftarrow \frac{1}{N} \);
   - Sample \((m, \Lambda)_{t+1}^{1:N} \sim K_t\);
4: end for
**Return:** Particles \((m, \Lambda)_T^{1:N}\).

Current discovery/exclusion method for reporting one or more possible mass values for the Higgs particle. The Bayes procedure can be calibrated to match specified frequency theory error rates.

### 3.1 Structure

The required ingredients of a decision theory problem are a model with the corresponding parameter space, a decision space which is a set of possible actions to take, and a loss function (Berger, 1980).

The model was introduced in Section 2. However, our procedure is introduced regardless of the specific details of the model. We define the decision space as the set of all possible subsets, \( S \subset M \). The interpretation of \( S \) is that, \( m \in S \) if, having observed the data, we wish to retain \( m \) as a possible value of the true mass. For instance, if \( 0 \in S \), the results suggest that it is possible that the Higgs particle does not exist (at least not with a mass in the search window). The loss function is defined in the following form,

\[
L(m, S) = \left\{ \left( C(m) + \int_S l(x)dx \right) 1_{S^c}(m) + \left( \int_S l(x)dx \right) 1_S(m) \right\} 1_{(x_0, x_n)}(m)
+ \left\{ \left( C(0) + \int_S l(x)dx \right) 1_{S^c}(0) + \left( \int_S l(x)dx - l(0) \right) 1_S(0) \right\} 1_{(0)}(m), \tag{5}
\]

where \( C(x) \) and \( l(x) \) are the exclusion and inclusion loss functions, respectively (giving the loss when mass values are excluded from or included in the decision set incorrectly), and \( S^c \) is the complement of \( S \).

By averaging the loss function (5) with respect to the marginal posterior \( \pi(m \mid y) \) the
posterior expected loss or the Bayes risk is obtained as follows.

\[
\begin{align*}
    r_{\pi(m|y)}(S) &= E_{\pi(m|y)}[L(m, S)] \\
    &= \int_{S^c} C(m)\pi(m | y)dm + \int_S l(m)dm \\
    &\quad + C(0)1_{S^c}(0)\pi(0 | y) - l(0)1_S(0) (1 - \pi(0 | y))
\end{align*}
\]

The Bayes rule is obtained by minimizing the Bayes risk with respect to \( S \);

**Theorem 1.** The Bayes rule, i.e., the decision rule that minimizes \( r_{\pi(m|y)}(S) \), is given by,

\[
S = \{ m; m \in M : \frac{l(m)}{C(m)} < \pi(m | y) \},
\]

Also,

\[
0 \in S \quad \text{if} \quad \frac{l(0)}{C(0)} < \frac{\pi(0 | y)}{1 - \pi(0 | y)},
\]

where \( C(0) \) and \( l(0) \) are the incorrect discovery and failure in discovery losses, respectively.

The proof is provided in Appendix A.

### 3.2 Calibration and error rate estimation

As mentioned before, the proposed procedure can be calibrated to satisfy frequency theory error rates. The loss ratios \( \frac{l(m)}{C(m)} \) can be adjusted to satisfy the type I error rates required in particle physics applications. The global type I error rate and false exclusion rates are controlled respectively as follows.

\[
\begin{align*}
    P(0 \notin S|m = 0) &= P\left( \frac{\pi(0 | y)}{1 - \pi(0 | y)} < \frac{l_0}{C_0} \right) \\
    &= \alpha_1,
\end{align*}
\]

\[
\begin{align*}
    P(m \notin S|m) &= P\left( \pi(m | y) < \frac{l(m)}{C(m)} \right) \\
    &= \alpha_2.
\end{align*}
\]

Solving the above equations for the loss ratios \( l_0/C_0 \) and \( l(m)/C(m) \) requires obtaining the \( \alpha_1100\% \) and \( \alpha_2100\% \) quantiles of the null distributions of the posterior odds \( \frac{\pi(0|y)}{1-\pi(0|y)} \), and the posterior probability, \( \pi(m|y) \).

Unfortunately, under most realistic models the distribution of the posterior functionals cannot be obtained in closed form. Recently, Johnson (2013) developed the uniformly most powerful Bayesian test (UMPBT) for one-parameter exponential family based on the same idea, i.e., maximizing the probability that the Bayes factor is smaller than a certain threshold.
under the null model. They briefly visit the Higgs problem and report the size of a Bayes factor equivalent to the local significance level of $\alpha_0 = 3 \times 10^{-7}$. However, to be able to obtain the UMPBT, a normal model is used and the cross section is treated as an unknown parameter as it is in the existing analysis of the Higgs data.

The results in Johnson (2013) cannot be used under our model. Therefore, we need to estimate percentiles of the distribution of the posterior using Monte Carlo. However, this requires intense computation since for each generated data set at each iteration of the Monte Carlo we need to run the SMC algorithm to estimate the posterior. This Monte Carlo within Monte Carlo scheme is computationally costly on its own, while satisfying the small significance level in the physics application requires a large number of iterations to estimate precise tail quantiles which makes the computation even more intense.

To address calibration with affordable computation, we combine importance sampling and approximation techniques: we replace the SMC algorithm with a Laplace approximation in each Monte Carlo algorithm and use importance sampling to reduce the number of iterations required to obtain tail probability estimates for the Bayesian statistic’s distribution for a fixed level of precision.

### 3.2.1 Approximation

In the following, we explain the Laplace approximation to the marginal posterior distribution of the mass of the Higgs particle. This approximation is used as a fast alternative sampling the posterior distribution to speed up the calibration Monte Carlo. The approximation method, inspired by Rue et al. (2009), is based on a Gaussian approximation to the conditional distribution, $\tilde{\pi}(\Lambda | m, y)$, i.e.,

$$
\pi(m \mid y) = \int \pi(\Lambda, m \mid y) d\Lambda \\
= \pi(m) \int \pi(\Lambda \mid m, y) d\Lambda \\
\approx \pi(m) \int \tilde{\pi}(\Lambda \mid m, y) d\Lambda.
$$

The Gaussian approximation, $\tilde{\pi}(\Lambda \mid y, m)$, is obtained by numerically approximating the mode and curvature of $\pi(\Lambda \mid y, m)$;

$$
\pi(\Lambda \mid y, m) \propto \exp\left\{-\frac{1}{2}(\Lambda - \mu)^T \Sigma^{-1}(\Lambda - \mu) + \log P(y \mid \Lambda, m)\right\}.
$$
Consider the Taylor expansion of the $n$ components of the log likelihood around the initial values $\Lambda_0$,

$$
\log P(y|\Lambda, m) = \sum_{i=1}^{n} g_i(\Lambda_i)
\approx \sum_{i=1}^{n} [g_i(\Lambda_{0i}) + g_i'(\Lambda_{0i})(\Lambda_i - \Lambda_{0i}) + \frac{g_i''(\Lambda_{0i})}{2} (\Lambda_i - \Lambda_{0i})^2]
= \sum_{i=1}^{n} [a(\Lambda_{0i}) + b_i(\Lambda_{0i})\Lambda_i - \frac{1}{2} c_i(\Lambda_{0i})\Lambda_i^2].
$$

Therefore, we have

$$
\tilde{\pi}(\Lambda | y, m) \propto \exp\{-\frac{1}{2}(\Lambda - \mu)^T\Sigma^{-1}(\Lambda - \mu)
+ \sum_{i=1}^{n} [a(\Lambda_{0i}) + b_i(\Lambda_{0i})\Lambda_i - \frac{1}{2} c_i(\Lambda_{0i})\Lambda_i^2]\}
= \exp\{-\frac{1}{2} \Lambda^T (\Sigma^{-1} + \text{diag}(c_0)) \Lambda + (\Sigma^{-1} \mu + b_0)^T \Lambda\}.
$$

where $b_0 = (b_1(\Lambda_{01}), \ldots, b_n(\Lambda_{0n}))^T$ and $c_0 = (c_1(\Lambda_{01}), \ldots, c_n(\Lambda_{0n}))^T$. The mean (mode) of the approximate Gaussian distribution, $\tilde{\pi}(\Lambda | y, m)$, is obtained by repeatedly solving $(\Sigma^{-1} + \text{diag}(c_0))\Lambda_{t+1} = (\Sigma^{-1} \mu + b_0)$ for $\Lambda_{t+1}$ until converged. The approximate covariance matrix is $\Sigma^{-1} + \text{diag}(c)$. Therefore, the approximate marginal distribution can be obtained up to a normalizing constant as follows,

$$
\tilde{\pi}(m | y) \propto \pi(m) \int \tilde{\pi}(\Lambda | m, y) d\Lambda 
\propto \pi(m)|\Sigma^{-1} + \text{diag}(c)|^{-\frac{1}{2}}.
$$

### 3.2.2 Importance sampling for estimating error rates

As mentioned before, to evaluate the error rates associated with the Bayesian testing procedure, tail probabilities of the posterior functionals need to be estimated. Accurate Monte Carlo estimates for probabilities of rare events are only obtained with large Monte Carlo samples (in the order of $10^7$ and larger in this application). In this section we introduce an importance Monte Carlo algorithm that is used to obtain tail probability estimates with lower variances.

We focus on the global type I error rate of the Bayesian testing procedure, i.e.,

$$
\alpha_1 = P_{H_0}(\frac{\pi_0(y)}{1 - \pi_0(y)} < \frac{l_0}{C_0}) = P(\pi_0(y) < q_0)
$$

where $\pi_0(y) = \pi(0 | y)$ and $q_0 = \frac{l_0}{1 - l_0/C_0}$. While in calibrating the Bayes procedure the goal is to estimate $q_0$ to satisfy a determined $\alpha_1$, suppose, for the time being, that $\alpha_1$ is to be estimated for a given $q_0$. 

10
To estimate $\alpha_1$ using basic Monte Carlo, data, $y_i$, is generated in each iteration under the null hypothesis, $H_0$, and $\pi_0(y_i)$ is obtained. The Monte Carlo estimate of $\alpha_1$ based on a (large) sample $N$ posterior values is given by,

$$\hat{\alpha}_1 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{(0,q_0)}(\pi_0(y_i))$$

However, under the null hypothesis, the event that $\pi_0$ falls below $q_0$ is rare and an unaffordably large $N$ is required to obtain a non-zero estimate for $\alpha_1$.

Importance sampling is a popular method for simulating rare events (Rubino and Tuffin, 2009). The idea is to generate samples under an importance distribution under which the event of interest is likely to occur and weight the samples according to the original distribution of interest. To use importance Monte Carlo, here, we seek an importance distribution under which small values of $\pi_0$ are more likely to occur.

Let us remind ourselves of the model under the null and alternative hypotheses,

- $H_0$: $m = 0$, $\Lambda \sim \pi(\Lambda)$, $y \sim P(y \mid \Lambda, m = 0)$
- $H_A$: $m \sim \pi_A(m); m > 0$, $\Lambda \sim \pi(\Lambda)$, $y \sim P(y \mid \Lambda, m)$

Clearly we expect the event $\pi_0(y) < q_0$ to occur with high probability under the alternative. Therefore we can use the model under $H_A$ as the importance distribution. The importance weights are then given by,

$$W_i = \frac{P(y \mid H_0)}{P(y \mid H_A)} = \frac{P(y \mid m = 0)}{\int_{x_0}^{x_n} P(y \mid m)dm} = \frac{\pi_0(y)P(y)/\pi_0}{\int_{x_0}^{x_n} [\pi(m \mid y)P(y)/\pi_A(m)]dm} = \frac{\pi_0(y)}{\pi_0 \int_{x_0}^{x_n} [\pi(m \mid y)/\pi_A(m)]dm}, \quad (8)$$

where $\pi_0$ is the prior distribution of $H_0$. The importance Monte Carlo estimate of $\alpha_1$ based on a sample generated under the alternative model is given by,

$$\tilde{\alpha}_1 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{(0,q_0)}(\pi_0(y_i))W_i. \quad (9)$$

For calibration, however, (9) is solved for $q_0$ with a given significance level, $\alpha_1$. Algorithm 2 outlines the calibration steps.
Algorithm 2 Importance Monte Carlo calibration algorithm

Input: Pre-determined significance level, $\alpha_1$.

1: for $i := 0, 1, \ldots, N$ do
   Generate $m_i \sim \pi_A(m)$;
   Generate $\Lambda_i \sim \pi(\Lambda)$;
   Generate data, $y_i \sim P(y \mid \Lambda_i, m_i)$;
   Obtain $\tilde{\pi}(m \mid y_i)$ using (6);
   Obtain $W_i$ using (8).
2: end for
3: Solve (9) to obtain $q_0$.

Return: Discovery threshold $q_0$.

4 The Higgs simulated data analysis

4.1 Data

Confidentiality rules in the area of particle physics do not allow access to the real data for non-members of the Higgs research groups. In this section we apply our model and procedure to simulated data provided to us by Matthew Kenzie of the CMS group and described in CMS Collaboration (2014) and CMS Collaboration (2013). The simulation procedure is very complex because it must model not only the predicted behaviour of the Higgs boson but also the behaviour of the extremely complex CMS detector. Analysis of such simulated data was an essential step in developing the analytic techniques to be used for the real experiment. The simulated data available to us represent the diphoton decay mode invariant mass spectrum (in the range $100 < m_{\gamma\gamma} < 180$ GeV) at centre of mass energy $\sqrt{s} = 8$ TeV. For each of the decay modes there are different Higgs signatures referred to as analysis categories. For the diphoton decay mode there are nine analysis categories. The simulated data to which we had access consist of a list of the invariant mass of each data event together with the corresponding analysis category.

There are several “production mechanisms” through which a Higgs particle can be generated; five such mechanisms were considered for our data. Each such production mechanism leads to a specific predicted signal function. While the production mechanism is not identified in the data the signal function is propagated through the analysis separately for each of the SM Higgs production mechanisms at the LHC and each analysis category. The shape of the signal function in each analysis category, for each of the production modes and at three hypothesized Higgs masses (120, 125, and 130 GeV) has been provided to us in form of a histogram; the entry in a single bin of such a histogram is the expected number of Higgs events produced by a specific production mechanism in a specific analysis category in the mass range for that bin if the Higgs has the particular hypothesized mass. A handful of the production mechanism, analysis category combinations produce so few expected out-
comes that we were not provided the corresponding signal histograms; in the end we have histograms for 41 of the 45 combinations.

We fit the signal function, \( (3) \), to these histograms and estimate the signal strength, \( c \), and signal width, \( \epsilon \). We then extrapolate from the three signal masses available to obtain the signal function corresponding to other masses. As a simplification we combine the data for all the nine analysis categories and bin the data according to the signal histograms with 322 bins. Figure 1 shows the histogram of the data. We also use the sum of the signal functions over the production modes and analysis categories as a single signal function for each mass.

Note that, in principle, our method can handle the complications produced by having several combinations of production mechanism/decay mode/analysis category. However, simulated data is provided to us for only one decay mode and in practice, low event counts in some of the analysis categories can be a source of problems in making inference separately in these categories.

### 4.2 Inference

Following the model introduced in Section 2, the background is modelled as a log-Gaussian process with known prior mean and covariance functions. The prior background mean func-
tion is chosen to be a fourth-order Bernstein polynomial function,

\[ \mu(z) = \sum_{k=0}^{4} b_i g_i(z), \]  

(10)

where \( z \in (0, 1) \) is the projection of mass \( (x \in (x_0, x_n)) \), into the unit interval to assist specification of the length scale parameter in the spatial covariance function, (2). The basis functions, \( g_i(z) \), are given by,

\[ g_i(z) = \binom{4}{i} z^i (1-z)^{4-i}, \]

and \( b_i \) are selected by fitting the polynomial to the data. While using the data to define the prior mean seems cyclical we note that the results are not sensitive to the mean function as long as the prior covariance structure of the log-Gaussian process allows large variability on the background curve. In the covariance function, (2), the variance parameter, \( \sigma^2 \), is chosen to provide the desired level of uncertainty (\( \sigma^2 = \log(5) \) which translates to average variance of about 10^5 under the log-Gaussian prior). The use of parametric models, such as fourth-order polynomials, for the background function in CMS Collaboration (2012) and ATLAS Collaboration (2012) reveals a strong prior belief in smoothness of the background. Therefore, we choose the length scale parameter, \( l \), to give a strong spatial correlation which would result in smooth sample paths under the prior.

Note that the choice of the hyper-parameters can influence the posterior distribution of \( m \); by choosing a larger value for \( l \) (assuming strong spatial correlation), occurrence of a rise of the type observed in the data at \( m \approx 126.5 \) becomes unlikely under the prior and the procedure would be more sensitive to such fluctuations in the data. Also, by choosing a smaller variance parameter, \( \sigma^2 \), one makes strong assumptions about how far the background curve could vary over the function space and data points falling far from the mean relative to the variance of the background model would be considered evidence for the possible location of the signal.

The prior distribution defined for the parameter \( m \) is a mixture distribution given by,

\[ \pi(m) = 0.5 \mathbb{1}_0(m) + 0.5 \mathcal{U}(m; 0, 1), \]

where \( \mathbb{1}_0 \) denotes a point mass at 0 and \( \mathcal{U} \) is the uniform density function.

The sampling step in Algorithm 1 is performed by generating the background and mass from proposal distributions followed by an accept/reject step. The proposal distribution for the background at time \( t+1 \) is a log-Gaussian distribution with mean equal to the background at time \( t \) and covariance matrix following the prior covariance structure but scaled according to the posterior variances at time \( t \). The proposal distribution for \( m \) is a kernel density estimate to the posterior of \( m \) at time \( t \). The temperature schedule, \( \{\tau_t\} \), is chosen to be a grid of size 50 on \([0, 1]\).

The results in terms of the posterior samples of mass and background curves over the 50 steps of the sequential algorithm are presented in Figure 2. The grey bands show the 95%
Figure 2: The results of the analysis of the simulated data representing the sequence of power posteriors through 50 steps of the sequential algorithm. The estimated mass of the Higgs particle ($m_H \approx 126$) is specified by the vertical red line.

credible intervals for the background function; the wide bright bands refer to the early steps of the algorithm where curves are sampled from the diffuse prior while as the credible bands get darker they become more focused around the target posterior mean. Underneath the background credible bands are kernel density estimates to the posterior distribution of mass, $m$, along the sequence of distributions. The colour of the curves get darker as the sample is trimmed and moved towards the target posterior. The red vertical line shows the mass value reported as the estimated mass of the Higgs particle ($m \approx 126$). The right hand side axis is scaled according to the posterior density of $m$.

The final results, i.e., the target posterior is plotted alone in Figure 3. This figure shows the posterior mean and 95\% credible intervals of the background as well as the posterior mean background curve plus the signal function centred at the maximum a posteriori (MAP) mass.

4.3 Calibrated discovery threshold and the decision set

We run Algorithm 2 for the model introduced in Section 2 to obtain the discovery threshold. Figure 4 shows the simulated posterior probabilities that the Higgs particle does not exist with their corresponding weights under the null hypothesis, i.e., the null distribution of the posterior probability, $\pi_0(y)$. The lower $\alpha_1 100\%$ quantile, $q_0 = 0.0066$, is identified by a
Figure 3: The results of the analysis of the simulated data representing the target posterior for the background and the posterior background mean plus the signal function centred at the maximum a posteriori mass. The estimated mass of the Higgs particle ($m_H \approx 126$) is specified by the vertical red line. The discovery decision based on the simulated data, $y_{sim}$, is made by comparing the estimated null posterior probability to the obtained threshold. The SMC-based point estimate of $\pi_0(y_{sim})$ is zero, i.e., no samples with $m = 0$ remain in the final posterior sample of size 5,000. A conservative 90% upper limit on this estimate would be about 0.0063 which is smaller than the discovery threshold. Therefore, we conclude that the simulated data contains adequate evidence of the existence of the Higgs particle.

Having concluded that $0 \notin S$, to obtain the final decision set we only need to obtain the exclusion thresholds, $q(m)$, for $m \in (100, 180)$, such that,

$$P(\pi(m \mid y) < q(m)) = \alpha_2,$$

where $\alpha_2 = 0.05$ is a common choice. Obtaining the exclusion thresholds is analogous to that of discovery threshold given in Section 3.2. Since the exclusion controlled error rates are not as small as the discovery thresholds even basic Monte Carlo can provide accurate estimates. However, given that the exclusion step is not as sensitive as the discovery step, running Monte Carlo for every mass (i.e., a fine discretization on the mass spectrum) seems to be waste of computation. Moreover, a Bayesian credible set for the mass of the Higgs particle can be obtained which is most likely very close to a decision set obtained by comparing
the exclusion thresholds to the posterior density for each mass. Therefore, we report the 95% Bayesian credible set, as a set of plausible mass values for the Higgs particle based on the posterior analysis of the simulated data. A 95% Bayesian credible set, $S_B$, is one that satisfies

$$\int_{S_B} \pi(m \mid y) = 0.95.$$ 

An estimate of $S_B$ is obtained as an interval whose lower and upper limits are given by the 0.025 and 0.975 sample quantiles of the posterior over mass: $\hat{S}_B = (124.26, 127.46)$.

## 5 Conclusion

We have proposed a Bayesian procedure that can be used as an alternative to the detection/exclusion method used in search for a new particle in particle physics. In a decision theoretic framework, we define a linear loss function that summarizes the possible outcomes of search for a new particle with the associated losses. The Bayes rule is obtained by minimizing the expected loss and is the basis of decision making.

We introduce a Bayesian hierarchical model to make Bayesian inference about the mass of the Higgs particle (the parameter of interest) as well as the background function. A
sequential Monte Carlo algorithm is used to obtain samples from the posterior distributions of the model parameters. The model is fit to data produced by computer models that simulate the behaviours of the detectors. The analysis results are consistent with the analysis of the real Higgs data reported in physics literature - the posterior distribution of the mass is peaked about the reported mass of the Higgs particle. The 95% credible Bayesian interval is reported as the set of credible values of the Higgs boson based on this analysis.

Our Bayes procedure can be calibrated to give required frequency theory error rates. A calibration algorithm is proposed in which the posterior distributions are obtained by a fast Laplace approximation instead of SMC thereby making the calibration step computationally feasible. An importance Monte Carlo for simulation of rare events is proposed that enables us to calibrate the Bayes procedure to meet the small type I error rate typically used in particle physics.

Acknowledgments

We are grateful to Louis Lyons and Matthew William Kenzie of the CMS group who provided to us the simulated data as well as insight on the problem throughout the course of this work.
A Proof of Theorem 1

Proof. Consider the problem of deciding whether or not mass value $m_0 \in M$ should be included in the decision set. Following the Bayes rule, $m_0$ is included if the risk associated with inclusion of $m_0$ is negative. First, consider the case that $m_0 \in (x_0, x_n)$. To avoid measure theoretic complications we restrict the decision space to finite unions of open intervals and $\{0\}$. Since the prior distribution of $m$, on the interval $(x_0, x_n)$ is absolutely continuous, the addition of any point like $m_0$ (as well as any other zero measure set) to the decision set leaves the risk unchanged. Let $S$ be the decision set before including $m_0$. We consider the change in the risk as a result of adding the interval $(m_0 - \delta, m_0 + \delta)$ for very small $\delta$ to $S$. Let $S_0 = S \cup (m_0 - \delta, m_0 + \delta)$. We include $m_0$ in the final decision set if and only if,

$$r(S_0) - r(S) < 0,$$

where

$$r(S_0) - r(S) = \int_S l(m)dm + \int_{m_0+\delta}^{m_0-\delta} l(m)dm + \int_{S^c} C(m)\pi(m | y)dm$$

$$- \int_{m_0-\delta}^{m_0+\delta} C(m)\pi(m | y)dm - \int_S l(m)dm - \int_{S^c} C(m)\pi(m | y)dm$$

$$= \{l(m_0) - C(m_0)\pi(m_0 | y)\} 2\delta + o(\delta).$$

The above expression is negative for all sufficiently small $\delta > 0$ if and only if

$$l(m_0) - C(m_0)\pi(m_0 | y) < 0,$$

or

$$\frac{l(m_0)}{C(m_0)} < \pi(m_0 | y).$$

For the case that $m_0 = 0$, let $S_0 = S \cup \{0\}$,

$$r(S_0) - r(S) = \int_S l(m)dm + l(0) (1 - \pi(0 | y)) + \int_{S^c} C(m)\pi(m | y)dm$$

$$- C(0)\pi(0 | y) - \int_S l(m)dm - \int_{S^c} C(m)\pi(m | y)dm$$

$$= l(0) (1 - \pi(0 | y)) - C(0)\pi(0 | y).$$

Therefore, $0$ is added to $S$ if

$$l(0) (1 - \pi(0 | y)) - C(0)\pi(0 | y) < 0$$

$$\frac{l(0)}{C(0)} < \frac{\pi(0 | y)}{1 - \pi(0 | y)}.$$
References

ATLAS Collaboration (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716:1–29.

Berger, J. O. (1980). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York Inc.

CMS Collaboration (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716:30–61.

CMS Collaboration (2013). Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel. Technical Report CMS-PAS-HIG-13-001, CERN, Geneva.

CMS Collaboration (2014). Observation of the diphoton decay of the Higgs boson and measurement of its properties. *Eur. Phys. J. C*, 74:3076.

Cowan, G., Cranmer, K., Gross, E., and Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *European Physical Journal C - Particles and Fields*, 71:1554–1573.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika*, 74:33–43.

Del Moral, P. D., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *J. R. Statist. Soc. B*, 68:411436.

Gross, E. and Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *European Physical Journal C - Particles and Fields*, 70:525–562.

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *The Annals of Statistics*, 41:1716–1741.

Rubino, G. and Tuffin, B. (2009). *Rare Event Simulation Using Monte Carlo Methods*. Wiley.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B*, 71:319–392.