A Frequency-Calibrated Bayesian Search for New Particles

Shirin Golchi*[1] and Richard Lockhart[1]

[1]Simon Fraser University, Department of Statistics and Actuarial Science

Abstract

The statistical procedure used in the search for new particles is investigated in this paper. The discovery of the Higgs particles is used to lay out the problem and the existing procedures. A Bayesian hierarchical model is proposed to address inference about the parameters of interest while incorporating uncertainty about the nuisance parameters into the model. In addition to inference, a decision making procedure is proposed. A loss function is introduced that mimics the important features of a discovery problem. Given the importance of controlling the "false discovery" and "missed detection" error rates in discovering new phenomena, the proposed procedure is calibrated to control for these error rates.

Keywords: Bayes rule, decision set, Higgs boson, linear loss function, sequential Monte Carlo

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 (Englert and Brout, 1964; Higgs, 1964; Guralnik et al., 1964; Guralnik, 2009). The existence of the Higgs boson was confirmed by experiments run at the Large Hadron Collider (LHC) at the European organization for nuclear research, known as CERN, in 2012 (CMS Collaboration, 2012; ATLAS Collaboration, 2012). LHC is a high energy collider specifically designed and constructed to detect the Higgs particle. Two beams of protons circulating at very high energies 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

*e-mail: sgolchi@sfu.ca

known SM particles and cannot 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 candidates". The "invariant mass" of the unobserved particle is computed from masses, energies and momenta of the decay products such that it reflects the conservation of mass-energy. A histogram of the estimator of the mass is then created for each decay mode (some analyses including $H \to 2$ photons, i.e. Higgs decays to two photons, use unbinned likelihood fits instead of histograms). 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 byproducts which pass the cuts. Effectively then, the SM predicts that if the Higgs particle exists and has mass m_H then there will be a bump on the histogram of invariant masses whose size and shape are completely predicted by unknown mass, m_H, and other measured quantities such as beam energy. The statistical problem is to determine whether or not such a bump exists and if so at what mass it is centered.

The search procedure has two stages. The first stage is discovery, or search, and the second is called exclusion. In the first stage, 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; this local p-value is then plotted against mass, typically on a log-scale, and the smallest value of this local value is reported as a quantity of significance (see Figure 7 and the discussion in Section 7 in (ATLAS Collaboration, 2012) and Figure 15 and related discussion in CMS Collaboration (2012)). At the same time a global p-value is computed by approximating the distribution of the smallest local p-value. A claim of "discovery" should require this global p-value to be less than $1 - \Phi(5)$ (a 5σ effect $-$ Φ is the standard normal cumulative density function). The second stage is referred to as exclusion where for each mass in the range under consideration the null hypothesis that the particle exists, at this mass, with the predicted signal strength, is tested against a background only hypothesis.

In the searches for the Higgs particle described in CMS Collaboration (2012) and ATLAS Collaboration (2012), in the discovery phase, theoretical predictions are used to define the signal function for different production mechanisms and analysis categories. An overall signal strength parameter in the form of a unitless scaling factor, generally denoted by μ, is used in the model. The standard model predicts that $\mu = 1$ and the null hypothesis that there is no Higgs is represented by $\mu = 0$. The procedures described in CMS Collaboration (2012) and ATLAS Collaboration (2012) treat the alternative to $\mu = 0$ as $\mu > 0$. Treating μ as a free parameter. It is worth noting that the signal strength parameter μ is the focus of the local likelihood ratio tests since the mass is fixed for each test and the background model parameters and any other unknown parameters of the signal function are treated as nuisance parameters. Consequently, simultaneous two dimensional confidence intervals for the mass of the Higgs particle and the signal strength parameter are reported for different decay modes.

In order to fix some notation we now give some further details of the two stages for the current procedure. Some key weaknesses of the existing method are highlighted in the process.

Search: For each mass value, $m \in (m_0, m_n)$, in the search window, a likelihood ratio test (see Cowan et al. (2011) for details) is performed for the hypotheses

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_A : \mu > 0. \tag{1}$$

Note that the overall signal strength, $\mu = 1$, is not used under the alternative. For each $m \in (m_0, m_n)$, a p-value is obtained. The smallest of these local p-values is described as the *local significance* and the mass at which this minimum is achieved is reported as an initial estimate of the mass of the Higgs particle. We note that this is not the final estimate of the mass; a further detailed analysis of the decay products in the relevant events is used to provide an estimate with uncertainties. The plot of local p-values does not itself provide an interval estimate for the mass of the detected particle.

This local significance corresponds to a testing procedure which rejects the null hypothesis if any of a family of local test statistics, $v(m)$, (e.g., log likelihood ratios) indexed by the unknown mass m is larger than a predefined level κ. 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:

$$\begin{aligned} P_G &= P_{H_0}(\max_m v(m) > \kappa) \\ &\approx P_{H_0}(\chi^2_d > \kappa) + \mathrm{E}_{H_0}\left(N(\kappa)\right) \end{aligned} \tag{2}$$

where κ in this case is the observed maximum statistic, χ^2_d, denoting a chi-squared distribution with d degrees of freedom, is the null distribution of $v(m)$ and $\mathrm{E}_{H_0}\left(N(\kappa)\right)$ is the expected number of upcrossings of the level κ by the process $v(m)$. 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,

$$\begin{aligned} \alpha_G &= P_{H_0}(\max_m v(m) > \kappa_{\alpha_0}) \\ &\approx P_{H_0}(\chi^2_d > \kappa_{\alpha_0}) + \mathrm{E}_{H_0}\left(N(\kappa_{\alpha_0})\right) \\ &= \alpha_0 + \mathrm{E}_{H_0}\left(N(\kappa_{\alpha_0})\right), \end{aligned} \tag{3}$$

where κ_{α_0} is the χ^2_d quantile corresponding to α_0. Since $\mathrm{E}_{H_0}(N(\kappa)) \geq 0$, 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 on the search range. For this very reason, discovery is announced if the global p-value shows a 5σ effect.

Exclusion: In the second stage, further investigation is done to exclude regions of m which are unlikely values for the mass of the Higgs boson. The theoretical signal strength is tested at significance level $\alpha_2 = 0.05$ at the exclusion step (Cowan et al., 2011),

$$H_0 : \mu = 1 \qquad \text{vs.} \qquad H_A : \mu < 1. \tag{4}$$

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

Some implications of the existing procedure are as follows; the two stages serve different purposes: the search stage is sensitive to a variety of signals rather than only that predicted by the SM, potentially searching for any Higgs-like phenomenon. The exclusion stage on the other hand is designed to exclude mass values at which a signal exactly matching the predicted signal is not observed. Different significance levels are used for the two steps which implies that much more caution is taken for testing for signal associated with a new phenomenon than for testing if an observed signal matches the predicted signal.

Some care is needed in interpreting the local and global significance levels particularly when both are presented in the same analysis. Use of a global 5σ rule for declaring 'discovery' controls the rate at which incorrect discoveries are announced provided that the rule is applied without any prior screening. To make this clear imagine the following reporting procedure: if the local significance level exceeds 5σ we compute and report the global p-value defined above. If the local significance level does not reach 5σ we neither compute nor report the global p-value. Then the interpretation of any global significance level exceeding 5σ is clear because whenever the global significance level exceeds 5σ so does the local significance level. Thus the global p-value will be computed in every case where it would exceed 5σ and in those cases bears the same interpretation as any small p-value. But when the global p-value is 4.5σ, say, we might or might not report it depending on the local significance level. Thus we cannot easily compute the probability that a reported global significance level will exceed 4.5σ.

It will be seen that in (ATLAS Collaboration, 2012) the word "discovery" is used once in the paper (in the conclusion section); the global significance for a search over the range 110 to 600 GeV is 5.1σ. The word 'discovery' is used in (CMS Collaboration, 2012) only in the introduction to describe the goal because the global significance, even over the narrower range of 115 to 130 Gev is only 4.6σ. (We note that both paper titles begin by "Observation of a new ..." and that except for the crucial word the discussions are quite similar.)

Both papers compute global p-values over more than one mass range, often increasing the global significance by searching over a smaller range. These smaller ranges are not without motivation; they often correspond to searching over mass ranges not excluded in earlier searches.

Data analysis does not end with discovery and exclusion. Having declared a discovery, for instance, confidence intervals for the mass and joint confidence sets for the signal strength parameter μ will be computed. The model will be assessed by checking to see if the signal strength is close to 1 and by checking to see if the μ's for different production mechanisms appear to be the same. In this paper we focus on the discovery/exclusion part of the problem in a partially Bayesian framework and provide Bayesian tools for this post-discovery analysis.

In the following we consider two strategies. First we propose a purely Bayesian approach where we treat the discovery of the Higgs particle as an inference problem. A Bayesian hierarchical model is defined that captures the main features of the particle discovery problem. Treating the mass of the Higgs boson as the main parameter of interest we estimate the background, signal strength parameter and the hyper-parameters in a fully Bayesian framework where uncertainties about the estimates are expressed by the posterior variance. A joint credible set for the mass and signal strength is proposed as the basis of discovery.

Our second strategy is decision theoretic. Given that a larger penalty is associated with false discovery than with missed detection, we propose a hybrid Bayes-frequentist decision making procedure that allows for controlling frequency error rates at desired levels while using a Bayesian test statistic such as the posterior odds. The proposed method can be perceived as a unified search procedure that takes advantage of the flexibility of the Bayesian framework while adhering to frequency theory requirements.

Our procedure is similar to that of Feldman and Cousins (1998), a generalization of which (Cowan et al., 2011) is used for construction of two dimensional confidence intervals for mass and signal strength. Feldman and Cousins (1998) discuss the coverage issues of confidence intervals that are built based on the two step procedure without proper conditioning and propose a method for constructing confidence intervals based on ranking the ratio of the likelihood under null and alternative models. Our method is similar to this approach in that the posterior odds are used for constructing the decision set. The final decision set is interpreted as a confidence interval with different confidence levels for values of the parameter (mass) that represent the null and the alternative hypotheses.

Note that systematic errors are not considered in the present work. However, such errors are naturally included in a Bayesian model as priors. More specifically, some systematic errors are difficult to deal with in a non-Bayesian framework. For instance, in computing the signal shape and cross section as a function of beam luminosity it is necessary to carry out quantum mechanical calculations by expansions where the result is an approximation error that typically can only be quantified subjectively. Incorporating that subjectivity explicitly in a prior offers the opportunity to examine the sensitivity of the conclusions to prior uncertainty.

The rest of the paper is organized as follows. In Section 2, we introduce a Bayesian hierarchical model and describe inference as a Bayesian search for the Higgs particle. The inference results are reported for simulated data provided by the CMS group. Section 3 is dedicated to the Bayesian decision making procedure that, while combining the discovery and exclusion steps, is calibrated to obtain desired frequency-theory error-rates. Approximation and sampling techniques are introduced for efficient implementation of calibration. Section 4 follows with concluding remarks and discussion of extensions and future work. In addition, a supplementary material document is available on-line that includes the proposed approach applied to a simple signal detection problem (the “on/off problem” – Section C).

2 Bayesian Inference

In this section, we propose a Bayesian hierarchical model that captures the features of the problem of discovery of the Higgs particle. The goal is to make inference about two parameters of interest, the mass of the Higgs particle and the signal strength parameter, while adequately handling the nuisance parameters that specify the background. The proposed model is fit to simulated data provided to us by the CMS group and the results are presented.

2.1 Model

Suppose that the data, i.e., the invariant masses recorded by the detector, are realizations of a Poisson process whose intensity function is given by the sum of a background intensity function $\Lambda(m)$ and a signal intensity function $s_{m_H}(m)$. The shape of the signal function is known and its location is determined by the unknown parameter, $m_H \in \mathcal{M}$, where $\mathcal{M} = \{\emptyset\} \cup (m_0, m_n)$ ($(m_0, m_n) \subset \mathcal{R}^+ - \{0\}$, i.e., m_0 and m_n are strictly positive). The parameter, m_H, is the unknown mass of the Higgs particle where $m_H \in (m_0, m_n)$ means that the Higgs boson has a mass in the search window, (m_0, m_n), while $m_H = \emptyset$ refers to the case that the particle does not exist, at least not with a mass in (m_0, m_n). Note that m_H is the parameter of interest and in the Bayesian framework is treated as a random variable; it should not be confused with m that is used to denote an arbitrary but fixed mass in $\mathcal{M}$.

We quantify uncertainty in the background by considering the logarithm of the intensity function to be a realization of a Gaussian process,

$$\log \Lambda_{\boldsymbol{\beta},\eta,\sigma^2}(m) \sim \mathcal{GP}(\xi_{\boldsymbol{\beta},\boldsymbol{\sigma^2}}(m), \rho_{\eta,\sigma_2}(m, m')), \quad m \in (m_0, m_n). \tag{5}$$

with covariance function,

$$\rho_{\eta,\sigma_2}(m, m') = \sigma^2 \exp(-\eta(m - m')^2), \tag{6}$$

where σ^2 is the variance parameter and η is the correlation parameter that controls the smoothness of the background function.

We parametrize the mean function, $\xi_{\boldsymbol{\beta}}(m)$, such that the expectation of the background, $E(\Lambda)$, has one of the typical parametric forms currently used to model the background function, for example a fourth order Bernstein polynomial (ATLAS Collaboration, 2012). Therefore,

$$\xi_{\boldsymbol{\beta},\sigma^2}(m) = \log\left(\sum_{i=0}^{4} \beta_i h_i(z(m))\right) - \frac{\sigma^2}{2}, \tag{7}$$

where $z : \mathcal{M} \rightarrow (0, 1)$ is an affine transformation of mass onto the unit interval and the basis functions are given by,

$$h_i(z) = \binom{4}{i} z^i (1 - z)^{4-i} \qquad z \in (0, 1). \tag{8}$$

However, we note that the choice of Bernstein polynomials is commonly made since they are constrained to be positive. Since under our model the log-Gaussian prior guarantees the positivity of background, an uncostrained parametric form may be used for the mean function.

The notation $\Lambda_{\boldsymbol{\beta},\eta,\sigma^2}$ is used to show the dependence of the background function on the hyper-parameters. For the sake of brevity the subscript is dropped from here on.

We choose the signal function as a Gaussian probability density function with the location parameter m_H (in the current practice (CMS Collaboration, 2012; ATLAS Collaboration, 2012) a slightly more complex signal shape called the "crystal ball function" is used). Thus, the signal function is given by

$$s_{m_H}(m) = c_{m_H}\,\phi\left(\frac{m - m_H}{\epsilon}\right) \qquad \text{for } m_H \in (m_0, m_n), \tag{9}$$

$$s_{\emptyset}(m) = 0, \tag{10}$$

where c_{m_H} is a scaling constant (analogous to the cross section), and ϕ is the standard normal probability density function. The standard deviation, ϵ, controls the spread of the signal function.

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

$$\pi(\mathbf{y}|m_H, \mu, \Lambda) = \prod_{i=1}^{n} \frac{\exp(-\Gamma_i)\Gamma_i^{y_i}}{y_i!}, \tag{11}$$

where

$$\Gamma_i = \int_{m_{i-1}}^{m_i} [\Lambda(m) + \mu s_{m_H}(m)]dm. \tag{12}$$

Here μ is the signal strength, i.e., a unitless parameter that allows for credibility of signal sizes that do not match the one predicted by the theory (equivalent to $\mu = 1$). The grid $\mathbf{m} = (m_0, m_1, \ldots, m_n)$ is the vector of bin boundaries over the search window. In practice, bin-wise expectations γ_i are distributed according to a multivariate normal distribution whose covariance matrix is obtained by applying (6) to the bin centers. This simplification is made to avoid integration of the log-Gaussian process at every evaluation of the likelihood for more efficient computations.

The posterior distribution of the model parameters $\boldsymbol{\theta} = (m_H, \mu, \Lambda, \boldsymbol{\beta}, \eta, \sigma^2)$ given the data $\mathbf{y}$ can be written as

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{13}$$

where

$$\pi(\boldsymbol{\theta}) = \pi(m_H)\pi(\mu)\pi(\Lambda \mid \boldsymbol{\beta}, \eta, \sigma^2)\pi(\boldsymbol{\beta})\pi(\eta)\pi(\sigma^2), \tag{14}$$

We describe the prior distribution on $\mathcal{M}$ by specifying a density with respect to the measure ν which puts a point mass on $\emptyset$ and normalized Lebesgue measure on (m_0, m_n).

That is,

$$\nu(A) = \mathbb{1}(\emptyset \in A) + \int_{A\cap(m_0,n_m)} \frac{dm}{m_n - m_0}, \tag{15}$$

where $\mathbb{1}(\cdot)$ is an indicator function. The prior on m_H is specified as a density, $\pi(m_H)$, with respect to ν given by

$$\pi(m_H) = 0.5 \tag{16}$$

This prior implies that a priori the two events, "Higgs particle does not exist" and "Higgs particle exists with mass $m_H \in (m_0, m_n)$" are equally likely. Then given that the Higgs particle exists, all the mass values in the search window are also equally likely. For clarity the prior probability of a subset A of $\mathcal{M}$ is

$$\Pi(A) \equiv \int \mathbb{1}(m \in A)\pi(m)\nu(dm) = \frac{1}{2}\mathbb{1}(\emptyset \in A) + \frac{1}{2}\int_{A\cap(m_0,m_n)} \frac{dm}{m_n - m_0}. \tag{17}$$

The choice of a particular prior for a real physical parameter like the Higgs mass will inevitably be controversial. Our frequency adjusted decision theoretic approach described in Section 3 will remove much of the impact of the prior; the point is discussed further there.

The signal strength parameter, μ, is assigned a log-normal prior with mean 1 and standard deviation of 0.6 which allows signals of about two tenth to about three times the size of the theoretical signal to remain credible under the prior. Note that the signal strength is meant for searching alternative theories to the SM. Therefore, the prior distribution needs to be specified in accordance with credible alternative signal sizes.

The background prior, $\pi(\Lambda \mid \boldsymbol{\beta}, \eta, \sigma^2)$ is given by (5) and the mean hyperparameters, $\boldsymbol{\beta}$, have diffuse normal priors. The correlation parameter, η, is assigned an inverse-Gamma prior with shape and scale parameters equal to one. This prior is considered weakly informative given the affine transformation of m onto $(0, 1)$ explained above. The prior over the variance parameter σ^2 is $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$ which is recommended in the literature as a reference prior (Paulo, 2005).

2.2 Computational issues

The mixture prior on m_H results in difficulties in Markov chain Monte Carlo (MCMC) sampling from (13). More specifically, moving efficiently between $\emptyset$ and (m_0, m_n) in the parameter space is not trivial which leads to poor mixing of the Markov chain. To overcome computational difficulties we use sequential Monte Carlo (SMC) (Doucet et al., 2001; 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. Samples are filtered through the sequence of densities using importance sampling to obtain a sample from the target distribution.

A common approach for defining the sequence of densities between the initial and target distributions is by inducing the likelihood in the model sequentially. Let the filtering sequence of distributions be denoted by,

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

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

$$\pi_t = \pi(\boldsymbol{\theta})[\pi(\mathbf{y}|\boldsymbol{\theta})]^{\tau_t}, \tag{18}$$

where

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

The SMC sampler comprises iterative steps of weighting and sampling. While the samples are moved toward the target distribution through re-sampling with weights calculated according to the current temperature, they are also moved toward higher probability regions under each distribution, using MCMC steps. These steps are explained in Algorithm 1.

The form of the incremental weights $\tilde{w}_i$ depends on the choice of transition kernel K_t in SMC. For a variety of choices for the forward kernel and importance weights in SMC see Del Moral et al. (2006). In Algorithm 1 K_t is chosen as an MCMC transition kernel that results in the simplified form of the incremental weights as the proportion of two consecutive densities evaluated for the current sample.

The proposal distributions used for the the MCMC step are as follows. For the background at time t a log-Gaussian distribution is used that has a mean equal to the background at time $t-1$ and covariance matrix following the prior covariance structure but scaled according to the posterior variances at time t. The proposal distribution for m is determined by estimating the marginal distribution of m given $\mathbf{y}$ at time $t-1$ as

$$\hat{\pi}_t(m \mid \mathbf{y}) = \begin{cases} \frac{\sum_{i=1}^{N} \mathbf{1}(m_t^i = \emptyset)}{N} & m = \emptyset \\ \frac{\sum_{i=1}^{N} \mathbf{1}(m_t^i \neq \emptyset)}{N} \zeta(m) & m \neq \emptyset. \end{cases} \tag{19}$$

where $\zeta(m)$ is a kernel density estimate of the distribution of $m \neq \emptyset$ at time t. The rationale for this proposal distribution is that for efficient Metropolis-Hastings steps, the proposal distribution needs to be chosen close to the target distribution. Construction of the proposal based on π_{t-1} is based on the assumption that π_{t-1} is close to π_t. A normal proposal is used for μ and the background mean hyper-parameters $\boldsymbol{\beta}$ and the covariance hyper-parameters are generated from chi-squared proposal distributions.

The temperature schedule, $\{\tau_t\}_{t=1}^{T}$, is specified adaptively using the approach proposed by Jasra et al. (2011): at each step t, τ_t is determined such that the effective sample size (ESS) is equal to a pre-specified value such as half the sample size. This is achieved by solving the following equation numerically for τ_t,

$$\text{ESS} = \frac{\left(\sum_{n=1}^{N} w_n^t(\tau_t)\right)^2}{\sum_{n=1}^{N} \left(w_n^t(\tau_t)\right)^2}. \tag{20}$$

The sampling is stopped when $\tau = 1$ satisfies the above condition.

Note that in simpler models, such as the one introduced for the on/off problem in Section C of the supplementary material, the above computational undertakings would not be necessary. The posterior in Section C is obtained by more direct numerical calculations.

Algorithm 1 Sequential Monte Carlo

```
Input: A temperature schedule $\{\tau_t, t = 0, \dots, T\}$
       A MCMC transition kernel $K_t$
Generate an initial sample $\boldsymbol{\theta}_0^{1:N} \sim \pi_0$;
$W_1^{1:N} \leftarrow \frac{1}{N}$;
for $t := 1, \dots, T-1$ do
    • $W_t^i \leftarrow W_{t-1}^i \frac{\tilde{w}_t^i}{\sum \tilde{w}_t^i}$ where $\tilde{w}_t^i = \pi(\mathbf{y}|\boldsymbol{\theta}_t^{(i)})^{\tau_t - \tau_{t-1}}$, $i = 1, \dots, N$;
    • Re-sample the particles $\boldsymbol{\theta}_t^{1:N}$ with importance weights $W_t^{1:N}$;
    • $W_t^{1:N} \leftarrow \frac{1}{N}$;
    • Sample $\boldsymbol{\theta}_{t+1}^{1:N} \sim K_t$;
end for
Return: Particles $\boldsymbol{\theta}_T^{1:N}$.
```

2.3 Data and Inference Results

Confidentiality rules do not allow access to the real data for non-members of the Higgs research groups. In this section we fit the proposed model 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 behavior of the Higgs boson but also the behavior 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. We had access to 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 the 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 outcomes that we were not provided the corresponding signal histograms; in the end we have histograms for 41 of the 45 combinations.

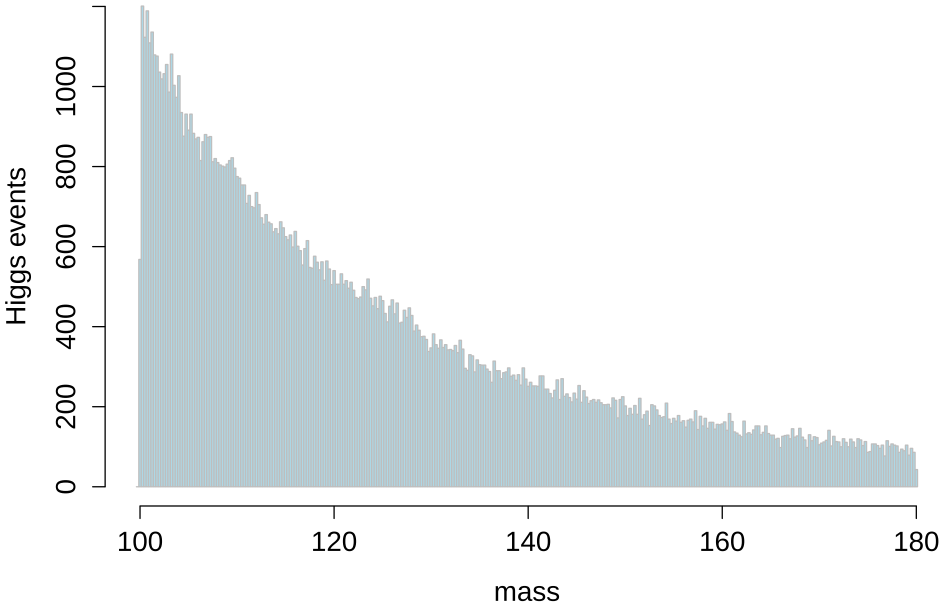


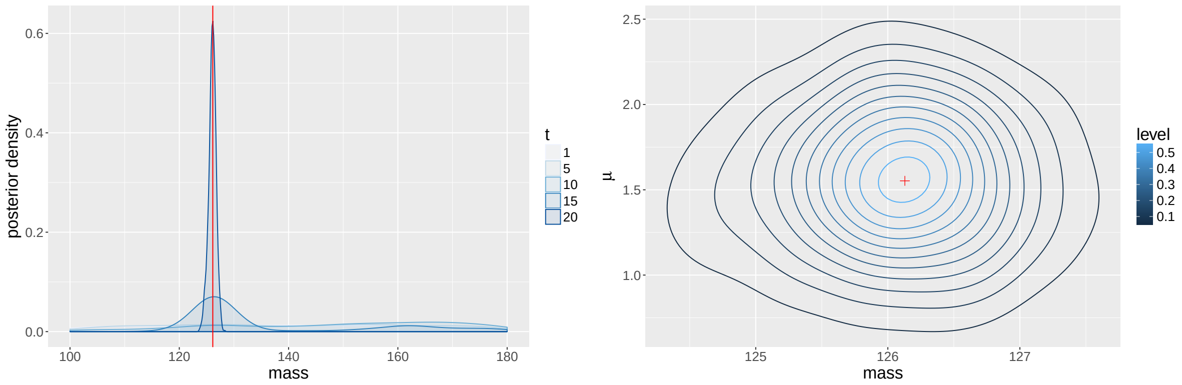
Figure 1: Simulated data representing the invariant masses of events

We fit the signal function, (10), to these histograms and estimate the signal strength, c_{m_H}, and signal width, ϵ. We then extrapolate from the three signal masses we were given to obtain the signal function corresponding to other masses. We pool the data for all nine analysis categories and bin the data according to the signal histograms with 322 bins. Figure 1 shows the histogram of the data. The reason for pooling the analysis categories is that some of these categories have very few data points and that the computational burden of a full analysis exceeded our capabilities. 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.

The results are summarized in Figures 2a, 2b and 3. Figure 2a shows the marginal posterior density estimates of mass over five selected steps of the sequential algorithm. The colour of the curves gets darker as the posterior gets closer to the target posterior. The red vertical line shows the maximum a posteriori mass value ($\hat{m}_H \approx 126$). A 95% credible interval for the mass of the Higgs particle based on the analysis of the simulated data is $(124.58, 127.30)$.

Figure 2b presents the joint posterior of mass and the signal strength parameter. The



(a) (b)

Figure 2: The results of the analysis of the simulated data representing (a) the posterior density of the mass evolving through the sequential sampler as the likelihood is induced into the power posterior sequentially; the estimated mass of the Higgs particle ($\hat{m}_H \approx 126$) is specified by the vertical red line, and (b) the joint posterior density contours of mass and signal strength; the red cross shows the position of mass and μ pair that maximizes the posterior.

contours are 2-d kernel density estimates of the joint posterior and the red cross shows the (m_H, μ) pair that maximizes the posterior.

In Figure 3 the blue bands show the 95% credible intervals for the background sample paths; the wide bands show the prior uncertainty for the background function. The background uncertainty decreases at each step of SMC as the data outweighs the prior. The background posterior mean and the background posterior mean plus the estimated signal are shown as a solid dark blue line and a dashed line, respectively.

3 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 current discovery/exclusion method for reporting one or more possible mass values for the Higgs particle. The Bayes procedure is calibrated to satisfy 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 2.1. However, the procedure we now suggest could be used regardless of the specific details of the model. We define the decision space as the

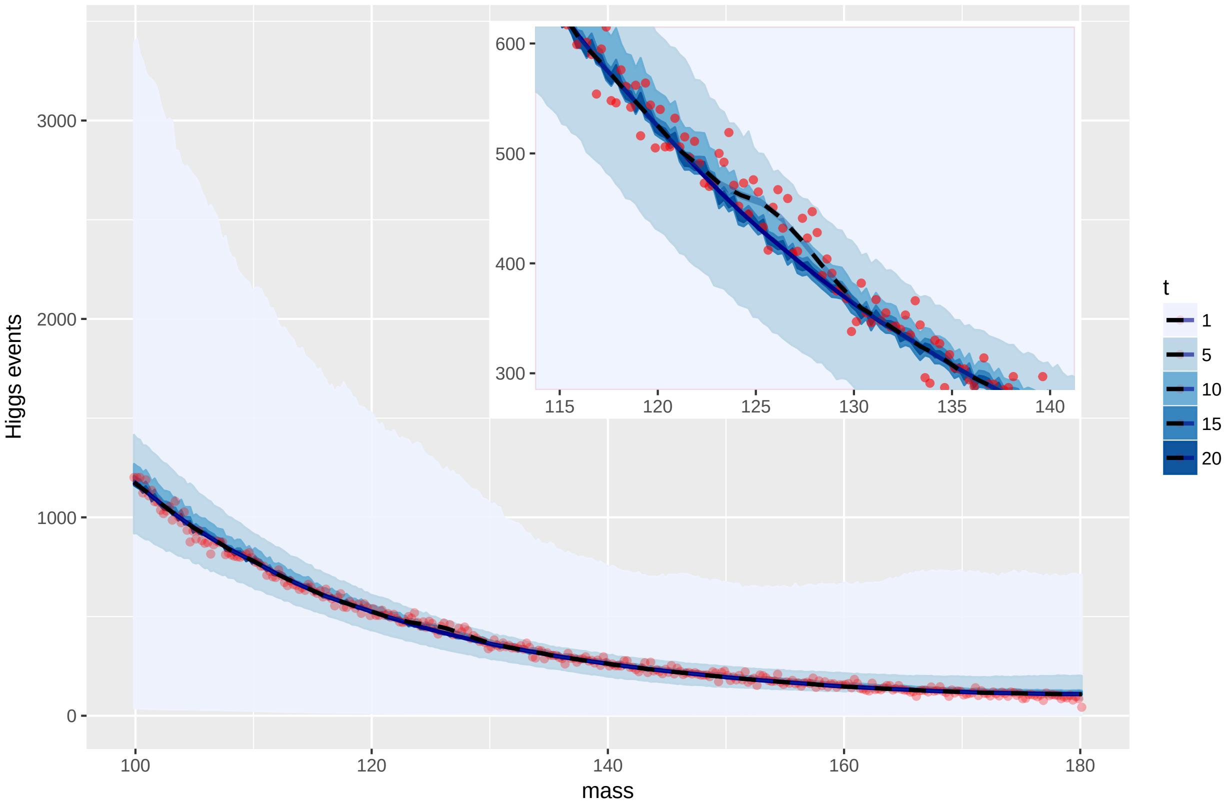


Figure 3: The results of the analysis of the simulated data representing the evolution of the 95% credible intervals for the background sample paths from the prior (light blue bands) to the posterior (dark blue bands). The red dots are the observations; the posterior background mean and the background posterior mean plus the estimated signal are shown by the solid dark blue line and the dashed line, respectively.

set of all possible subsets, $S \subset \mathcal{M}$ where $\mathcal{M}$ was defined in Section 2.1 as the union of the interval (m_0, m_n) and the single point $\emptyset$ that represents the case that the Higgs particle does not exist. 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 $\emptyset \in S$, 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 next step is to define a loss function that reflects two goals. First we would like to include the correct parameter value in the decision set. Therefore we charge a penalty if the correct value is excluded from the decision set. Second, we would like to exclude from S any incorrect parameter value. So we charge a penalty for including any parameter value that is not the true value. For the time being, suppose that the parameter space, $\mathcal{M}$, and the decision set, S, are discrete. Let $\ell_i(m_j)$ and $\ell_e(m_j)$ denote the loss values that respectively correspond to the case where m_j is not the mass of the Higgs particle ($m_j \neq m_H$) but included in the decision set and the case where m_j is the true mass of the Higgs particle

$(m_j = m_H)$ but excluded from the decision set. We refer to $\ell_i(m_j)$ and $\ell_e(m_j)$ as inclusion and exclusion losses respectively. Allowing ℓ_i and ℓ_e to depend on m_j permits us later on to adjust these functions to give desired error rates.

The following linear loss function accounts for all the possible decision scenarios with the corresponding losses,

$$L_D(m_H, S) = \sum_{m_j \in \mathcal{M}} \ell_i(m_j)\mathbb{1}(m_j \in S)\mathbb{1}(m_j \neq m_H) + \ell_e(m_H)\mathbb{1}(m_H \notin S) \tag{21}$$

where the subscript D in L_D shows the momentary discreteness assumption. However, we wish to treat the mass as a continuous variable over the interval of interest. Moreover, the case $m_H = \emptyset$ needs to be displayed explicitly in the formulas since it is treated differently than the rest of the parameter space in terms of error rates. Therefore, we begin by rewriting the loss function as

$$\begin{aligned} L_D(m_H, S) = \Big[&\sum_{m_j \in S \cap (m_0, m_n)} \ell_i(m_j)\mathbb{1}(m_j \neq m_H) \\ &+ \ell_i^{\emptyset}\mathbb{1}(\emptyset \in S) + \ell_e(m_H)\mathbb{1}(m_H \notin S)\Big]\mathbb{1}(m_H \in (m_0, m_n)) \\ +\Big[&\sum_{m_j \in S \cap (m_0, m_n)} \ell_i(m_j)\mathbb{1}(m_j \neq m_H) + \ell_e^{\emptyset}\mathbb{1}(\emptyset \notin S)\Big]\mathbb{1}(m_H = \emptyset) \end{aligned} \tag{22}$$

To pass to the continuous case we now replace the sum over $S \cap (m_0, m_n)$ by an integral with respect to the measure ν defined in (15). To do so we replace $\mathbb{1}(m_j \neq m_H)$ by $\mathbb{1}(m_H \notin (m_{j-1}, m_j)$ and write

$$\sum_{m_j \in S \cap (m_0, m_n)} \ell_i(m_j)\mathbb{1}\left(m_H \notin (m_{j-1}, m_j)\right) = \sum_{m_j \in S \cap (m_0, m_n)} \frac{\ell_i(m_j)(m_n - m_0)}{m_j - m_{j-1}} \frac{\delta m_j}{m_n - m_0} \tag{23}$$

where $\delta m_j = m_j - m_{j-1}$. Then we take a limit as the largest δm_i converges to 0 and require the rescaled loss

$$\frac{\ell_i(m_j)(m_n - m_0)}{m_j - m_{j-1}} \tag{24}$$

to have a limit as $n \to \infty$ (where n represents the number of masses) and $m_j \to m$ which we still denote by $\ell_i(m)$; this abuse of notation should cause no confusion. The new function ℓ_i is a "loss density" with respect to normalized Lebesgue measure on (m_0, m_n); we also denote $\ell_i(\emptyset) = \ell_i^{\emptyset}$ for compact notation. Values of the function ℓ_i have units of "loss" as do the quantities $\ell_e^{\emptyset}$ and $\ell_e(m_H)$. Note that the exclusion loss is only defined for $m_i = m_H$. The result is the loss function

$$
\begin{aligned}
L(m_H, S) &= \int_S \ell_i(m)\mathbb{1}(m \neq m_H)\nu(dm) \\
&\qquad + \ell_e(m_H)\mathbb{1}(m_H \notin S)\mathbb{1}(m_H \in (m_0, m_n)) + \ell_e^{\emptyset}\mathbb{1}(\emptyset \notin S)\mathbb{1}(m_H = \emptyset) \\
&= \Big[\int_{S\cap(m_0,m_n)} \ell_i(m)\mathbb{1}(m \neq m_H)\frac{dm}{m_n - m_0} \\
&\qquad + \ell_i^{\emptyset}\mathbb{1}(\emptyset \in S) + \ell_e(m_H)\mathbb{1}(m_H \notin S)\Big]\mathbb{1}(m_H \in (m_0, m_n)) \\
&\quad + \Big[\int_{S\cap(m_0,m_n)} \ell_i(m)\mathbb{1}(m \neq m_H)\frac{dm}{m_n - m_0} + \ell_e^{\emptyset}\mathbb{1}(\emptyset \notin S)\Big]\mathbb{1}(m_H = \emptyset). \qquad (25)
\end{aligned}
$$

The indicator $\mathbb{1}(m \neq m_H)$ can be dropped from the first integral without changing its value to give the simplified form

$$
\begin{aligned}
L(m_H, S) = &\int_{S\cap(m_0,m_n)} \ell_i(m)\frac{dm}{m_n - m_0} + \left[\ell_i^{\emptyset}\mathbb{1}(\emptyset \in S) + \ell_e(m_H)\mathbb{1}(m_H \notin S)\right]\mathbb{1}(m_H \in (m_0, m_n)) \\
&+ \ell_e^{\emptyset}\mathbb{1}(\emptyset \notin S)\mathbb{1}(m_H = \emptyset), \qquad (26)
\end{aligned}
$$

where the term $\int_{S\cap(m_0,m_n)} \ell_i(m) d\frac{m}{m_n - m_0}$ is the loss due to including incorrect mass values in S.

By averaging the loss function (26) with respect to the marginal posterior $\pi(m \mid \mathbf{y})$ the posterior expected loss or the Bayes risk is obtained as follows.

$$
\begin{aligned}
r_{\pi(m|\mathbf{y})}(S) &= E_{\pi(m|\mathbf{y})}[L(m, S)] \\
&= \int_{S\cap(m_0,m_n)} \ell_i(m)\frac{dm}{m_n - m_0} + \int_{S^c\cap(m_0,m_n)} \ell_e(m)\pi(m \mid \mathbf{y})\frac{dm}{m_n - m_0} \\
&\quad + \ell_e^{\emptyset}\mathbb{1}_{S^c}(\emptyset)\pi(\emptyset \mid \mathbf{y}) + \ell_i^{\emptyset}\mathbb{1}_S(\emptyset)\left(1 - \pi(\emptyset \mid \mathbf{y})\right). \qquad (27)
\end{aligned}
$$

In (27) the terms in (26) were rearranged to make it easier to identify the Bayes procedure. 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 $r_{\pi(m|\mathbf{y})}(S)$, is given by*

$$
S = \begin{cases} \{m \in (m_0, m_n) : \frac{\ell_i(m)}{\ell_e(m)} < \pi(m \mid \mathbf{y})\} & \frac{\ell_i^{\emptyset}}{\ell_e^{\emptyset}} \geq \frac{\pi(\emptyset|\mathbf{y})}{1-\pi(\emptyset|\mathbf{y})} \\ \{m \in (m_0, m_n) : \frac{\ell_i(m)}{\ell_e(m)} < \pi(m \mid \mathbf{y})\} \cup \{\emptyset\} & \frac{\ell_i^{\emptyset}}{\ell_e^{\emptyset}} < \frac{\pi(\emptyset|\mathbf{y})}{1-\pi(\emptyset|\mathbf{y})}. \end{cases} \qquad (28)
$$

A proof is provided in Section A of the supplementary material.

By the above theorem, the optimal Bayes decision set is obtained by including all the values of $m \in (m_0, m_n)$ whose posterior density is greater than the ratio of inclusion loss to exclusion loss. The $\emptyset$ is included if the corresponding posterior odds ratio is greater than its inclusion to exclusion loss ratio. The loss ratios can be specified such that the Bayes decision set has certain frequency coverage. In the following we explain the calibration process.

3.2 Calibration and error rate estimation

As mentioned before, the proposed procedure is calibrated to give desired frequency theory properties such as error rates. The loss ratios $\ell_i(m)/\ell_e(m)$ can be adjusted to satisfy the type I error rates required in particle physics applications. The same effect could be achieved in principle by keeping the loss ratio fixed and adjusting the prior. However, fixing the prior makes the computations below more straightforward. The global Type I error rate and false exclusion rates are controlled by solving the following equations for $\ell_i(\emptyset)/\ell_e(\emptyset)$ and $\ell_i(m)/\ell_e(m)$, respectively.

$$\begin{aligned} P(\emptyset \notin S \mid m_H = \emptyset) &= P\left(\frac{\pi(\emptyset \mid \mathbf{y})}{1 - \pi(\emptyset \mid \mathbf{y})} \leq \frac{\ell_i^{\emptyset}}{\ell_e^{\emptyset}} \mid m_H = \emptyset\right) \\ &= \alpha_1, \end{aligned} \tag{29}$$

$$\begin{aligned} P(m \notin S \mid m_H = m, \mu = 1) &= P\left(\pi(m \mid \mathbf{y}) < \frac{\ell_i(m)}{\ell_e(m)} \mid m_H = m, \mu = 1\right) \\ &= \alpha_2, \end{aligned} \tag{30}$$

for $m \in (m_0, m_n)$ and where $P(A)$ is the probability of event A.

The Associate Editor has observed that these calibrations are not strictly frequentist in nature because of the way we are handling the background Λ. A strict frequentist procedure would regard the function Λ as a parameter. Thus such a frequentist procedure would have

$$P(\emptyset \notin \hat{S} | m_H = \emptyset, \Lambda) = \alpha_1, \tag{31}$$

$$P(m \notin \hat{S} | m_H = m, \Lambda, \mu = 1) = \alpha_2. \tag{32}$$

where $\hat{S}$ is used instead of S to indicate the dependence on data. Such calibration is likely impossible if we insist on exact calibration in small samples. Traditional statistical procedures such as likelihood ratio tests are calibrated by letting the critical value for the test statistic depend on the nuisance parameters. In other words, equation (31) is required to hold only at the estimated value $\hat{\Lambda}$ of Λ; approximate calibration is then achieved by parametric bootstrapping. The procedure we describe below is parallel but averages over those Λ which remain credible after seeing the data, i.e., according to the posterior distribution of Λ; details are given in the next two subsections. Our procedure and parametric bootstrapping are both properly calibrated in large samples.

In (30) and (32) exclusion is done for the Standard Model, that is, only at $\mu = 1$. In (32) parametric bootstrapping can be used with Λ replaced by an estimate. As for (29), standard large sample results predict that this will achieve calibration at the true but unknown background.

Solving equations (29) and (30) for the loss ratios $\ell_i^{\emptyset}/\ell_e^{\emptyset}$ and $\ell_i(m)/\ell_e(m)$ requires obtaining the $\alpha_1 100\%$ and $\alpha_2 100\%$ quantiles of the distributions of the posterior odds $\pi(\emptyset|\mathbf{y})/(1 - \pi(\emptyset|\mathbf{y}))$, and the posterior probability density, $\pi(m|\mathbf{y})$ under the null hypotheses for the two tests, i.e., $m_H = \emptyset$ and $m_H = m$, respectively.

Unfortunately, under most realistic models the distribution of the posterior functionals cannot be obtained in closed form. In Johnson (2013) the uniformly most powerful Bayesian test (UMPBT) for one-parameter exponential family is developed based on the same idea, i.e., maximizing the probability that the Bayes factor is smaller than a certain threshold under the null model. Johnson (2013) briefly visits the Higgs problem and reports 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.

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 adding to the computational intensity.

Note that calibration to produce desired frequency error rates offsets the effect of the prior in decision making. Since the posterior odds can be written as a product of the prior odds and marginal likelihood ratio, the loss ratio we obtain and use in decision making, responds to the choice of the prior. Nevertheless the prior must be expected to influence the inferences and certainly plays a role in the computational costs of the procedure. In particular the prior of Section 2 is computationally expensive and we have not implemented our calibration in this context as a result. Instead we replaced the conditional prior on μ, given $m_H = m$ for a mass $m \in (m_0, m_n)$ by a point mass at $\mu = 1$.

This choice was not popular with reviewers of this paper. Here is some discussion of the issue from a frequency theory perspective. For a frequentist the performance of a procedure $\hat{S}$ is evaluated by using the frequentist risk. This is the function of the unknown parameters m_H, μ, and Λ given by

$$R(m_H, \mu, \Lambda; \hat{S}) = \mathrm{E}\left\{L(m_H, \hat{S}) \mid m_H, \mu, \Lambda)\right\}. \tag{33}$$

This risk can be written as a sum of two components, $R_\emptyset$ and R_E, given by

$$R_\emptyset(m_H, \mu, \Lambda) = \left\{\ell_i^\emptyset \mathbb{1}(m_H \neq \emptyset) + \ell_e^\emptyset \mathbb{1}(m_H = \emptyset)\right\} P(\emptyset \in \hat{S} \mid m_H, \mu, \Lambda), \tag{34}$$

and

$$\begin{aligned} R_E(m_H, \mu, \Lambda) = &\int_{m_0}^{m_n} \ell_i(m) P(m \in \hat{S} | m_H, \mu, \Lambda) \frac{dm}{m_n - m_0} \\ &+ \ell_e(m_H) 1(m_H \neq \emptyset) P(m_H \in \hat{S} \mid m_H, \mu, \Lambda). \end{aligned} \tag{35}$$

The component $R_\emptyset$ depends only on the random set $\hat{S} \cap \{\emptyset\}$ while the second depends only on $\hat{S} \cap (m_0, m_n)$. Any such random subset of $\{\emptyset\}$ has a natural interpretation as a hypothesis test which rejects the null hypothesis of no Higgs if and only if the random subset is empty. The level of this test is

$$P(\emptyset \notin \hat{S} \mid m_H = \emptyset, \mu, \Lambda), \tag{36}$$

and the power at $m \neq \emptyset$, μ and Λ is

$$\mathrm{Pow}(m, \mu, \Lambda) = P(\emptyset \notin S \mid m_H = m, \mu, \Lambda). \tag{37}$$

Since the frequentist risk is additive our procedure minimizes

$$R_\emptyset(m_H, \mu, \Lambda; \hat{S}) \tag{38}$$

calibrated as at (29), i.e., it minimizes the (Bayesian) expected risk averaged over Λ subject to our constraint. It can also be thought of as incorporating a hypothesis test of $H_0 : m_H = \emptyset$. The level of the test is

$$\alpha_1 = P(\emptyset \notin \hat{S} \mid m_H = \emptyset, \Lambda), \tag{39}$$

and the power is the function

$$\mathrm{Pow}(m, \mu, \Lambda) = P(\emptyset \notin \hat{S} \mid m_H = m, \mu, \Lambda), \tag{40}$$

for $m \in (m_0, m_n)$. Using a prior on μ which puts a point mass on $\mu = 1$ will maximize $\mathrm{Pow}(m, 1, \Lambda)$ (averaged over Λ). The prior of Section 2 will maximize $\mathrm{Pow}(m, \mu, \Lambda)$ averaged over μ and Λ. Our reviewers argue that we should not limit our attention to $\mu = 1$ but we observe that this amounts to trading sensitivity at the theory being considered, the Standard Model, against sensitivity at theories which make similar predictions to those of the Standard Model but with $\mu \neq 1$.

The issue is important so we take the time to consider some elementary analogues. In a problem testing a one-sided hypothesis about a normal mean the most powerful level α test does not depend on the particular alternative; there is a uniformly most powerful test. After calibration any Bayesian procedure analogous to ours will then simply give back this uniformly most powerful test. In more complex problems the most powerful test, produced by the Neyman-Pearson Lemma, will depend on the alternative. For each given prior on the alternative we would get a different calibrated Bayesian-frequentist test which maximized the power, averaged with respect to that prior. This sort of problem is our situation. The procedure we present below makes the choice to use the highly informative point mass prior on μ largely for computational reasons. To highlight the computational intensity for the case that μ is allowed to vary, consider that the SMC algorithm when tun in serial to generate 1000 samples takes about two hours. Calibration requires simulating the posterior multiple times to obtain the distribution of the test statistic. Even 100 simulations would take about 8 days which needs to be repeated for each pair of mass and μ to obtain the exclusion threshold. (The time complexity is $O(N \times M \times G)$, where N is the number of posterior samples, M is the number of Monte Carlo simulations and G is the number of (m, μ) pairs.) Of course, by parallelizing over a large number of processors, full calibration can be feasible. However, such computational undertaking is beyond the scope of the present work. These computational considerations also make it very challenging to carry out a full power study comparing a variety of priors on the signal strength parameter.

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. Again, for the on/off problem (Section C of the supplementary material) approximation is not required as the posterior can be obtained by simpler numerical calculations.

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 to sampling the posterior distribution to speed up the calibration Monte Carlo. The hyper-parameters, $\boldsymbol{\beta}$, η and σ^2 are held fixed at their maximum a posteriori estimates in the calibration. Consider reparametrizing the model in terms of $\Psi = \log \Lambda$. The approximation method, inspired by Rue et al. (2009), is based on a Gaussian approximation to the conditional distribution, $\tilde{\pi}(\Psi \mid m, \mathbf{y})$, i.e.,

$$\tilde{\pi}(m \mid \mathbf{y}) = \left.\frac{\pi(m, \Psi \mid \mathbf{y})}{\tilde{\pi}(\Psi \mid m, \mathbf{y})}\right|_{\Psi=\Psi^*}. \tag{41}$$

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

$$\pi(\Psi \mid \mathbf{y}, m) \propto \exp\{-\frac{1}{2}(\Psi - \boldsymbol{\xi})^T \Sigma^{-1} (\Psi - \boldsymbol{\xi}) + \log \pi(\mathbf{y}|\Psi, m)\}, \tag{42}$$

where $\boldsymbol{\xi}$ and Σ are the mean vector and covariance matrix, respectively. Consider the Taylor expansion of the n components of the log likelihood around the initial values Ψ_0,

$$\begin{aligned}
\log \pi(\mathbf{y}|\Psi, m) &= \sum_{i=1}^{n} g_i(\Psi_i) \\
&\approx \sum_{i=1}^{n} [g_i(\Psi_{0i}) + g_i'(\Psi_{0i})(\Psi_i - \Psi_{0i}) + \frac{g_i''(\Psi_{0i})}{2}(\Psi_i - \Psi_{0i})^2] \\
&= \sum_{i=1}^{n} [a_i(\Psi_{0i}) + b_i(\Psi_{0i})\Psi_i - \frac{1}{2}c_i(\Psi_{0i})\Psi_i^2].
\end{aligned} \tag{43}$$

where,

$$g_i(\Psi) = \log(\pi(y_i \mid \Psi, m)), \tag{44}$$

$$a_i(\Psi_{0i}) = g_i(\Psi_{0i}) - g_i'(\Psi_{0i})\Psi_{0i} + \frac{g_i''(\Psi_{0i})}{2}\Psi_{0i}^2, \tag{45}$$

$$b_i(\Psi_{0i}) = g_i'(\Psi_{0i}) - \Psi_{0i} g_i''(\Psi_{0i}), \tag{46}$$

$$c_i(\Psi_{0i}) = -\frac{g_i''(\Psi_{0i})}{2}. \tag{47}$$

The above expressions are given explicitly in Section B of the supplementary material. Note that g_i and its derivatives and consequently (44 - 47) depend on m but the dependence is not explicitly expressed for the sake of conciseness. Therefore, we have

$$\begin{aligned}\tilde{\pi}(\Psi \mid \mathbf{y}, m) &\propto \exp\{-\frac{1}{2}(\Psi - \boldsymbol{\xi})^T\Sigma^{-1}(\Psi - \boldsymbol{\xi}) \\ &\qquad + \sum_{i=1}^{n}[a_i(\Psi_{0i}) + b_i(\Psi_{0i})\Psi_i - \frac{1}{2}c_i(\Psi_{0i})\Psi_i^2]\} \\ &\propto \exp\{-\frac{1}{2}\Psi^T(\Sigma^{-1} + \mathrm{diag}(\mathbf{c}_0))\Psi + (\Sigma^{-1}\boldsymbol{\xi} + \mathbf{b}_0)^T\Psi\}. \qquad (48)\end{aligned}$$

where $\mathbf{b}_0 = (b_1(\Psi_{01}), \ldots, b_n(\Psi_{0n}))^T$ and $\mathbf{c}_0 = (c_1(\Psi_{01}), \ldots, c_n(\Psi_{0n}))^T$. The mean (mode) of the approximate Gaussian distribution, $\tilde{\pi}(\Psi \mid \mathbf{y}, m)$, is obtained by repeatedly solving $(\Sigma^{-1} + \mathrm{diag}(\mathbf{c}_t))\Psi_{t+1} = (\Sigma^{-1}\boldsymbol{\xi} + \mathbf{b}_t)$ for Ψ_{t+1} until convergence, where $\mathbf{b}_t$ and $\mathbf{c}_t$ are updated matrices at iteration t. The approximate covariance matrix of Ψ is $\Sigma^{-1} + \mathrm{diag}(\mathbf{c})$, where $\mathbf{c}$ is the convergent value of the sequence $\mathbf{c}_t$. Therefore, the approximate marginal distribution can be obtained up to a normalizing constant as follows,

$$\begin{aligned}\tilde{\pi}(m \mid \mathbf{y}) &\propto \pi(m)\int \tilde{\pi}(\Psi \mid m, \mathbf{y})d\Psi \\ &\propto \pi(m)|\Sigma^{-1} + \mathrm{diag}(\mathbf{c})|^{-\frac{1}{2}}. \qquad (49)\end{aligned}$$

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.,

$$\alpha_1 = P\left(\frac{\pi(\emptyset \mid \mathbf{y})}{1 - \pi(\emptyset \mid \mathbf{y})} < \frac{\ell_i^{\emptyset}}{\ell_e^{\emptyset}} \mid m_H = \emptyset\right) = P(\pi(\emptyset \mid \mathbf{y}) < q_{\emptyset} \mid m_H = \emptyset) \qquad (50)$$

where $q_{\emptyset} = \ell_i^{\emptyset}/(\ell_e^{\emptyset} + \ell_i^{\emptyset})$. While in calibrating the Bayes procedure the goal is to estimate $q_{\emptyset}$ to satisfy a determined α_1, suppose, for the time being, that α_1 is to be estimated for a given $q_{\emptyset}$.

To estimate α_1 using basic Monte Carlo, data, $\mathbf{y}_i$, is generated in each iteration under the null hypothesis, $m_H = \emptyset$, and $\pi(\emptyset \mid \mathbf{y}_i)$ is obtained. The Monte Carlo estimate of α_1 based on a (large) sample of N posterior values is given by,

$$\hat{\alpha}_1 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\pi(\emptyset \mid \mathbf{y}_i) \in (0, q_{\emptyset})) \qquad (51)$$

However, under the null hypothesis, the event that $\pi(\emptyset \mid \mathbf{y}_i)$ falls bellow q_0 is rare and an unaffordably large N is required to obtain a non-zero estimate for α_1.

Importance sampling is a popular method for simulating rare events (Rubino and Tuffin, 2009). The idea is to generate samples from an importance distribution under which the event of interest is more 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 $\pi(\emptyset \mid \mathbf{y})$ are more likely to occur.

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

$$H_0 : \text{The Higgs particle does not exist, i.e., } m_H = \emptyset, \tag{52}$$

$$H_A : \text{The Higgs particle exists with a mass } m_H \in (m_0, m_n), \tag{53}$$

Clearly we expect the event $\pi(\emptyset \mid \mathbf{y}) < q_\emptyset$ 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,

$$\begin{aligned} W_i &= \frac{\pi(\mathbf{y} \mid H_0)}{\pi(\mathbf{y} \mid H_A)} \\ &= \frac{\pi(\mathbf{y} \mid m = \emptyset)}{\int_{m_0}^{m_n} \pi(\mathbf{y} \mid m) dm} \\ &= \frac{\pi(\emptyset \mid \mathbf{y})\pi(\mathbf{y})/\pi(\emptyset)}{\int_{m_0}^{m_n} [\pi(m \mid \mathbf{y})\pi(\mathbf{y})/\pi_A(m)] dm} \\ &= \frac{\pi(\emptyset \mid \mathbf{y})}{\pi(\emptyset) \int_{m_0}^{m_n} [\pi(m \mid \mathbf{y})/\pi_A(m)] dm}, \end{aligned} \tag{54}$$

where $\pi(\emptyset)$ and $\pi_A(m)$ are the priors over the mass under H_0 and H_A, respectively. The importance Monte Carlo estimate of α_1 based on a sample generated under the alternative model is given by,

$$\tilde{\alpha}_1 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\pi(\emptyset \mid \mathbf{y}_i) \in (0, q_\emptyset)) W_i. \tag{55}$$

For calibration, however, (55) is solved for $q_\emptyset$ with a given significance level, α_1. Algorithm 2 outlines the calibration steps. As mentioned earlier, the marginal posterior of the mass is obtained by integrating the background over its posterior distribution, i.e., the procedure is calibrated for the most likely realizations of the background function. This is done by repeatedly sampling from the posterior sample of the background generated by Algorithm 1 in step 1-b of Algorithm 2.

3.3 Calibrated discovery threshold and the decision set

We run Algorithm 2 for the model introduced in Section 2.1 with $\mu = 1$ to obtain the discovery threshold. The lower $(3 \times 10^{-7})100\%$ quantile of the null distribution of the posterior probability, $\pi(\emptyset \mid \mathbf{y})$ is $q_0 = 0.0066$. The discovery decision based on the simulated

Algorithm 2 Importance Monte Carlo calibration algorithm

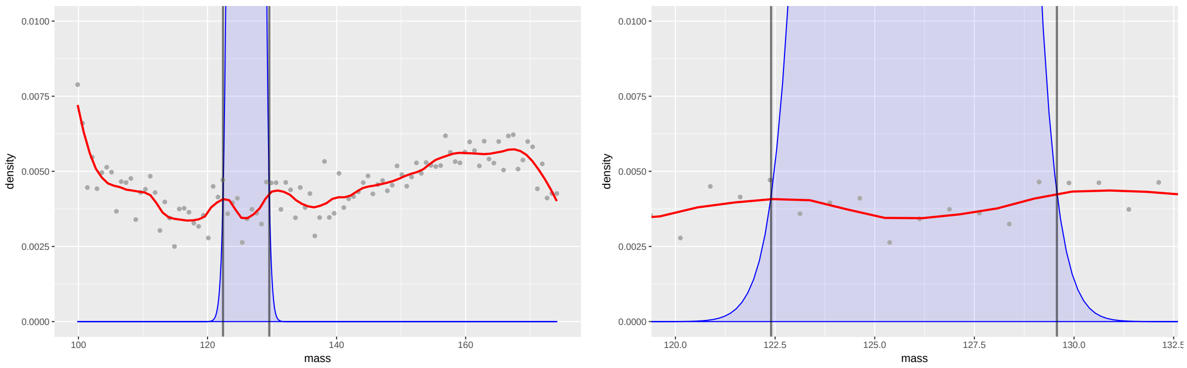
```
Input: Pre-determined significance level, $\alpha_1$.
for $i := 0, 1, \ldots, N$ do
    a- Generate $m_i \sim \pi_A(m)$;
    b- Generate $\boldsymbol{\Lambda}_i \sim \pi(\boldsymbol{\Lambda} \mid \mathbf{y}_{obs})$ (sample a realization from the posterior sample generated by Algorithm 1);
    c- Generate data, $\mathbf{y}_i \sim \pi(\mathbf{y} \mid \boldsymbol{\Lambda}_i, m_i)$;
    d- Obtain $\tilde{\pi}(m \mid \mathbf{y}_i)$ using (49);
    e- Obtain $W_i$ using (54).
end for
Solve (55) to obtain $q_0$.
Return: Discovery threshold $q_0$.
```

data, $\mathbf{y}_{sim}$ is made by comparing the estimated null posterior probability to the obtained threshold. The SMC-based point estimate of $\pi(\emptyset \mid \mathbf{y}_{sim})$ is zero, i.e., no samples with $m = \emptyset$ remain in the final posterior sample. Therefore, we conclude that the simulated data contains adequate evidence of the existence of the Higgs particle and we do not include $m = \emptyset$ in the decision set.

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

$$P(\pi(m \mid \mathbf{y}) < q(m) \mid m_H = m) = \alpha_2, \tag{56}$$

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, the approximation in (41) is pointwise and the computational burden increases with the size of the mass grid. Therefore, the time complexity of the exclusion step is $O(N^2 \times M)$ for a mass grid of size N and M Monte Carlo iterations. To reduce the computational cost we use a coarse discretization of the mass spectrum and use a kernel smooth of the exclusion thresholds for this coarse grid. Figure 4 shows the exclusion threshold together with the histogram of the posterior sample of mass in the search window. The boundaries of the decision set are determined by the mass values where the estimated posterior density is higher than the exclusion threshold. These boundaries are specified in Figure 4 by the grey vertical lines. The decision set obtained by the exclusion threshold cuts is $\hat{S} = (122.40, 129.57)$ which is more conservative than the 95% Bayesian credible set reported in Section 2.3, i.e., $(124.58, 127.30)$.



(a)

(b)

Figure 4: Exclusion threshold curve (red line) over the marginal posterior density estimate of mass together with the decision interval (the interval within two gray vertical lines). In (b) the graph is zoomed in for better visualization, the points are the decision thresholds obtained for the coarse grid over mass.

4 Conclusion and Discussion

In this article a Bayesian hierarchical model is introduced to make inference about the mass and signal strength of the Higgs particle as well as estimating the background function and its hyper-parameters in a fully Bayesian framework. A sequential Monte Carlo algorithm is used to sample 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 resemble the analysis of the real Higgs data reported in physics literature - the posterior distribution of the mass is peaked around the reported mass of the Higgs particle, i.e., $\hat{m}_H \approx 126$ (CMS Collaboration, 2012; ATLAS Collaboration, 2012). The 95% credible Bayesian interval is reported as the set of credible values of the Higgs boson based on this analysis.

In addition, we have proposed a frequency-calibrated Bayesian procedure that can be used as an alternative to the detection/exclusion method used in the search for the Higgs particle. 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.

Our procedure is 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 allows for calibration of the Bayes procedure to meet the small type I error rate typically used in discovery of new phenomena. We calibrate our procedure according to the error rate requirements in particle physics. A decision set is reported that excludes the null hypothesis (the hypothesis of “no Higgs”) and contains a range of masses that remain plausible according to discovery and exclusion error rates of 3×10^{-7} and 0.05 respectively.

As mentioned earlier, the nature of our procedure resembles that of Feldman and Cousins (1998) for constructing confidence intervals that discuss the coverage issues of confidence intervals built based on the two step procedure without proper conditioning. However, we use the posterior odds ratios instead of likelihood ratios and report a confidence sets with different levels of confidence for parameter values that represent the null hypothesis than those under the alternative hypothesis.

We conclude with a number of issues deserving acknowledgment and/or further work: There are a number of issues surrounding the inclusion of μ as an unknown parameter to be estimated. In Section 2 we have described a fully Bayesian analysis including μ. There will inevitably be controversy surrounding any choice of prior; we believe that the prior need to captures the physicists' consensus view of likely signal strength parameter values. Prior specification should, of course, involve a sensitivity analysis to study the robustness of the results to prior choice. Such sensitivity analysis is out of scope of the present paper but can be the focus of further investigations.

Another issue with μ is the possibility of including it in our decision theoretic approach. The decision theoretic procedure of Section 3 may be extended to consider the parameter space

$$\mathcal{M}_2 := \{\emptyset\} \cup \{(m, \mu) \mid m_0 < m < m_n, \mu > 0\} \tag{57}$$

to give a set $S_2 \subset \mathcal{M}_2$ which would be a variable level confidence set for the pair (m_H, μ). The losses would be extended to such pairs and we could try to calibrate more error rates than just those at (29) and (30). The resulting calibrations would be computationally much more demanding and would often yield a complex set of pairs (m_H, μ) for μ near 0. Preliminary testing has shown that calibration will require a lot more computation and likely some care in implementation; as a result we have not yet pursued this problem.

The set S of Section 3 combines discovery and exclusion in the sense that a value of m is excluded if it is not in S. The new set S_2 can be interpreted to do exclusion if we declare m to be excluded provided $(m, 1) \notin S_2$. The key is that we are excluding the Standard Model value of μ.

The Bayes procedure will be quite sensitive to the prior for μ; the decision theoretic procedure should be less so because calibration of the error rates reduces the impact of the prior. It does not eliminate the effect since the prior determines an ordering on the possible data sets which is used to determine which data sets correspond to a set S_2 including a given pair (m, μ).

We remark that a likelihood ratio statistic is a maximum over mass values of a family of test statistics depending on a mass m. This maximization produces the well known irregularities in the large sample theory of the test. Our proposal effectively uses the Bayes factor as a test statistic; this statistic replaces maximization with averaging, a much more regular process. We expect this regularization will make our calibration procedures work better with our test statistic than they might with a likelihood ratio.

A power study of the procedure in Section 3, comparing the effects of priors should be carried out. Such a study would require significant computational resources.

After the set S has been found further data analysis will be needed. If $\emptyset \notin S$, a discovery will be declared and it will be necessary to continue the analysis. One needs to:

- Give a suitable confidence set or credible region for m_H;
- Give a suitable confidence set or credible region for the pair (m_H, μ);
- Assess the assertion that $\mu = 1$ – If $\mu \neq 1$ then we might be discovering a particle which is Higgs-like but not the specific object predicted by the Standard Model; alternatively we might be discovering that systematic errors have not been adequately handled;
- Assess the assertion that μ does not depend on the production mechanism of the particle in question.

Our procedure and the standard procedure could, in principle, have discovered more than one peak had these been present, in the sense that our confidence set might find two or more disjoint intervals in (m_0, m_n).

Interpretation of such an occurrence is properly part of the post discovery analysis described above if we do not want that interpretation to contribute, potentially, to error rates in the formal discovery declaration step.

Finally, systematic errors were ignored in our discussion. Those errors are naturally included in a Bayesian model as priors. For systematic errors that result from measurement errors for parameters in the calculations leading to the mass histogram, one might imagine non-Bayesian (frequentist) approaches. But there are systematic errors which are much harder to deal with in frequentist terms. For instance, in computing the signal shape and cross section as a function of beam luminosity it is necessary to carry out quantum mechanical calculations by expansions. These expansions are carried out at some order described by phrases such as Next to Leading Order (NLO) or Next to Next to Leading order (NNLO). The result is an approximation error which typically can only be quantified subjectively. Incorporating that subjectivity explicitly in a prior offers the opportunity to examine the sensitivity of the conclusions to the specification of this prior uncertainty.

An adaptation of the proposed approach is presented for a signal detection problem, referred to as the on/off problem, in the supplementary material (Section C). As mentioned before, due to simplicity of the problem and model, the proposed approach can be implemented with less computational cost in this framework.

Data and Code

The data and code required for reproducing the results in the paper are provided at
https://github.com/sgolchi/BPD.

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.

We also greatly appreciate the very substantial time and effort spent by the associate editor and the reviewers to help improve the manuscript.

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):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(1):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.

Doucet, A., De Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice.* Springer-Verlag.

Englert, F. and Brout, R. (1964). Broken Symmetry and the Mass of Gauge Vector Mesons. *Physical Review Letters*, 13:321–323.

Feldman, G. J. and Cousins, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, 57:3873–3889.

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.

Guralnik, G. S. (2009). The History of the Guralnik, Hagen and Kibble Development of the Theory of Spontaneous Symmetry Breaking and Gauge Particles. *International Journal of Modern Physics A*, 24:2601–2627.

Guralnik, G. S., Hagen, C. R., and Kibble, T. W. B. (1964). Global conservation laws and massless particles. *Phys. Rev. Lett.*, 13:585–587.

Higgs, P. W. (1964). Broken Symmetries and the Masses of Gauge Bosons. *Physical Review Letters*, 13:508–509.

Jasra, A., Stephens, D. A., and Doucet, A. (2011). Inference for Lévy-Driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38:1–22.

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

Paulo, R. (2005). Default priors for gaussian processes. *The Annals of Statistics*, 33:556–582.

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.