Received: 15 February 2017 | Revised: 5 December 2017 | Accepted: 27 December 2017 | Published on: 19 January 2018

DOI: 10.1002/sam.11371

RESEARCH ARTICLE

WILEY

Informative priors in Bayesian inference and computation

Shirin Golchi

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada

Correspondence
Shirin Golchi, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada.
Email: golchi.shirin@gmail.com

Funding Information
National Science Foundation, SES-1023176. Natural Sciences and Engineering Research Council of Canada.

The use of prior distributions is often a controversial topic in Bayesian inference. Informative priors are often avoided at all costs. However, when prior information is available, informative priors are appropriate means of introducing this information into the model. Furthermore, informative priors, when used properly and creatively, can provide solutions to computational issues and improve inference. Through 3 examples with different applications, we demonstrate the importance and utilities of informative priors in incorporating external information into the model and overcoming computational difficulties.

KEYWORDS

aggregated relational data, constraints, differential equation models, particle discovery, sequential Monte Carlo, uncertainty

1 | INTRODUCTION

In most scientific problems, information is available from multiple sources and in different forms. While observational or experimental data is the basis of statistical modeling, it is rarely the case that data are the only information available about the unknowns. Various forms of information include the range of plausible values for the model parameters, information about the data generative process, and adherence assumptions of the data to some underlying deterministic function. In many cases, partial data may be available from multiple sources, such as previous studies or expert knowledge about a subset of unknown parameters. In classical statistics, such information is incorporated into the model in the form of (hard) constraints. However, hard constraints do not allow uncertainty about the available information and often result in inference and computational challenges.

In the Bayesian framework, constraints are often formulated into prior distributions. From a philosophical point of view, in Bayesian statistics, any additional knowledge besides data is considered information that can be used to improve inference rather than imposing restrictions. Stark [20] argues that the common Bayesian approach for incorporating hard constraints is not equivalent to imposing the constraint as is done in frequentist approaches. Considering the constraint to be a set of plausible values for the model parameters, Stark [20] argues that even a flat prior on this set is a much stronger assumption than the frequentist alternative that assumes no more than mere membership—a flat prior assumes that all values in the constraint set are equally likely. In a study by Gelman [9], similar arguments are made about imposing hard constraints using "non-informative" priors that can result in misleading the posterior distribution. Another example of relevant work in this area is that of Gelman et al. [11] who propose weakly informative default priors for regression models. They emphasize that their proposed method is different than the existing approaches in that, instead of fully informative priors that are application-specific or noninformative priors based on invariance principle, they use "a somewhat informative prior distribution that can nonetheless be used in a wide range of applications" and is interpreted as a generic constraint over the regression coefficients.

While the prior distribution is a natural and convenient means of incorporating information into the statistical model, it is one of the major reasons why Bayesian inference is often criticized [10]. The critics argue that if the researchers have the liberty of incorporating their subjective view into the statistical inference through the prior distribution, the objectivity of scientific research may be questioned.

Much discussion exists both in favor of and against Bayesian inference. See, for example, Bayarri and Berger [3], Little [14], and Gelman [10]. In this paper, however, we do not focus on the controversy around the use of prior distributions in Bayesian inference. The goal of this article is to highlight the role of prior distribution as a means of adequately using information that would otherwise be challenging to introduce

into a statistical model. We focus on problems with modeling constraints and external information. More specifically, we consider scenarios where hard constraints need to be replaced by soft constraints due to either prior uncertainty or methodological/computational challenges created by the constraints. Through 3 different examples, we showcase the use of informative priors to formally introduce uncertainty about the prior knowledge or temporarily relax constraints for computational purposes.

Our first example (Section 2) is focused on a Bayesian hierarchical model proposed by Zheng et al. [21] for analyzing aggregated relational data. External information about a subset of model parameters is required to identify the model. We demonstrate that informative priors can be used to realistically incorporate external information about some of the model parameters to overcome nonidentifiability, thereby addressing efficient computation and accessibility to the users while providing interpretable results.

In our second example (Section 3), we consider the Bayesian analysis of data used to discover the Higgs boson. We review the Bayesian hierarchical model proposed by Golchi and Lockhart [13] and illustrate the role of the prior used for the nuisance parameters in providing statistical power to detect the signal.

In the third example (Section 4), we consider inference for a system of ordinary differential equations (ODE) where adherence to the ODE is a hard constraint that creates computational difficulties. We emphasize that the use of a prior distribution in this example serves a different purpose compared to the first 2 examples. The prior is not a part of the model but is used as a computational maneuver. By introducing a prior distribution that allows temporary departure from the ODE model, one can overcome computational challenges. The results, however, remain unaffected as the relaxation of constraints is a computational hack to eventually fit the model with hard constraints. Section 5 follows with concluding remarks.

2 | ANALYSIS OF AGGREGATED RELATIONAL DATA

Aggregated relational data (ARD) are collected through social surveys by asking the respondents about the number of their acquaintances in various subpopulations. These subpopulations are, of course, selected according to the research goals. However, to facilitate statistical inference, a number of groups need to be specified as "auxiliary subpopulations" with known demographic information, for example, male and female names. While most ARD are collected to infer social structure, this type of data can be perceived as a summary of full network data (nodes and edges). Therefore, ARD are relevant in other areas of science where network data appear. Full network data is transformed into ARD by aggregating the links of a set of sample nodes over specific communities and ignoring the community membership of the sample nodes. While such aggregation results in the loss of information, it may be necessary in specific contexts, for instance, to preserve privacy.

The data, y_{ik}, comprises the number of acquaintances of individual $i = 1, \ldots, I$ in group $k = 1, \ldots, K$. The goal is to recover as much information as possible about the underlying network. To this end, the following model was proposed by Zheng et al. [21]:

$$y_{ik} \sim \text{Poisson}(\lambda_{ik}), \tag{1}$$

with $\lambda_{ik} = \gamma_{ik} \exp(\alpha_i + \beta_k)$, where α_i is the degree parameter; $\exp(\alpha_i)$ is the network size of respondent i; β_k is the prevalence parameter; $\exp\beta_k$ is the relative "sociability" of group k; and γ_{ik} is the individual propensity parameter measuring the tendency of respondent i to make ties with group k. It is correctly stated in Zheng et al. [21] that this model is mathematically unidentifiable as the likelihood depends on the 3 sets of parameters through a product form. In other words, one cannot learn about all the parameters of this model by relying on the data alone. Therefore, more information needs to be incorporated, through assumptions or external knowledge, to make inference possible.

Zheng et al. [21] propose to resolve the nonidentifiability issue in 2 steps; first, the model is simplified by integrating the likelihood with respect to the propensity parameters over a Gamma prior distribution with mean 1 and variance

$$var(\gamma_{ik}) = \frac{\omega_k - 1}{\exp(\alpha_i + \beta_k)}, \tag{2}$$

where ω_k represents the overdispersion associated with group k. The prior distribution of γ_{ik} implies the assumption that propensity parameters are a priori distributed around 1 but are less variable if the corresponding degree and prevalence parameters are larger. Integrating the likelihood with respect to γ_{ik} over this prior distribution results in the following negative binomial likelihood:

$$y_{ik} \sim \mathcal{NB}(\text{mean} = \exp(\alpha_i + \beta_k), \quad var = \omega_k \exp(\alpha_i + \beta_k)). \tag{3}$$

This simplification reduces the number of parameters by $K(I-1)$, where K is the number of groups, and I is the number of respondents. However, this does not resolve nonidentifiability entirely as the likelihood depends on α_i and β_k through their sum and remains unchanged for an infinite number of α and β pairs. At this stage, the auxiliary subpopulations with available information from external sources, such as the census, play a key role. Zheng et al. [21] use this information through a renormalization step embedded in the Markov chain Monte Carlo sampling. More specifically, the known subpopulation sizes are used to calculate a constant that is subtracted from the β_ks and added to the α_is, practically leading the Markov chain to the "correct whereabouts" of the parameter space.

The dataset used in Zheng et al. [21] is that of McCarty et al. [15], which was collated by asking $I = 1375$ individuals

the number of their acquaintances in each of the $K = 32$ subpopulations. The set of subpopulations comprises 12 male and female names and 20 groups of interest, such as the homeless, diabetic patients, gun dealers, etc. The 12 names are used as auxiliary subpopulations as the number of people with a certain name in the US population can be estimated from other sources, such as the census. The renormalization constant is specified as a linear function of the percentage of rare female names, somewhat popular female names, and somewhat popular male names in the US population. Zheng et al. [21] state that the renormalization procedure "is designed for the recall problems that exist in the McCarty et al. [15] dataset. Researchers working with different datasets may need to develop a procedure appropriate to their specific data."

The renormalization approach has a number of drawbacks: incorporating a portion of the external information into the model in this manner seems arbitrary and ad hoc; the expression of the renormalizing constant is data-specific and cannot be generalized to other ARD analysis. The authors have also not provided justifications that can be used as guidelines for the defining of such a constant in a different problem. In addition, hard-coding the renormalization step within the MCMC sampling is a road block for implementing the model in available MCMC softwares such as JAGS and Stan, which makes the model inaccessible for practitioners.

We argue that, as an alternative approach, external information should be used formally as a part of the model. In the Bayesian framework, the prior distribution is an appropriate way to introduce additional knowledge into the model and has the advantage of allowing uncertainty about this information. In the following sections, we explain an alternative model in which we use informative priors to incorporate existing information and assumptions into the model to overcome nonidentifiability. Moreover, we propose to use the available information for all the auxiliary groups rather than a subset of them. The resulting model is fit using Hamiltonian Monte Carlo implemented in Stan (http://mc-stan.org/) and is, therefore, readily accessible for practitioners.

Consider the original Poisson model in 1. The main advantage of this model is interpretability of the model parameters: the individual propensity parameters can be used as a relative distance metric between the respondents and subpopulations. Following Zheng et al. [21], we use normal prior distributions for the prevalence parameters β_k. However, we incorporate the additional information about the male and female names as the mean of the normal priors for the corresponding β_k. Small variances are chosen for these normal priors that, ideally, would be determined by uncertainty estimates of the provided information. In the current problem, however, no uncertainty estimates are given for the percentage of names in the US population. Therefore, these estimates are treated as "known" group proportions. The variances are selected as the smallest values that allow the proper mixing of the Markov chain. For the rest of β_k, we use diffuse normal priors centered at the mean of the given group sizes:

$$\beta_k \sim N\left(\log\left(\frac{N_k}{N}\right), \sigma_\beta^2\right), \quad k = 1, \ldots, K_1, \tag{4}$$

where $\frac{N_k}{N}$ is given, and σ_β^2 are small ($\sigma_\beta^2 = 0.01$);

$$\beta_k \sim N\left(\frac{1}{K_1}\sum_{k=1}^{K_1}\log\left(\frac{N_k}{N}\right), \tau_\beta^2\right), \quad k = K_1{+}1, \ldots, K, \tag{5}$$

where τ_β^2 are chosen to allow large deviations from the mean ($\tau_\beta^2 = 100$).

As in Zheng et al. [21], the prior distribution for α_i is:

$$\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2), \tag{6}$$

where μ_α and σ_α are assigned noninformative priors—$\mathcal{N}$ (0,100) and $\mathcal{IG}$ (0.001, 0.001), respectively.

For the propensity parameters γ_{ik}, we define a prior which assumes that individuals are equally likely to form ties with all groups ($E[\gamma_{ik} = 1]$) with enough flexibility that would let the data decide otherwise. The following gamma distribution represents this assumption:

$$\gamma_{ik} \sim \Gamma(a_\gamma, b_\gamma), \tag{7}$$

where $a_\gamma = b_\gamma$ are chosen such that the prior variance is relatively large. We choose $a_\gamma = b_\gamma = 0.25$ that is equivalent to a variance of 4.

With the choice of priors, the parameters of this hierarchical model can be estimated using Stan. The Stan implementation of the model can be accessed at https://github.com/sgolchi/ARDmodels.

A brief sensitivity analysis suggests that increasing σ_β^2 up to a certain threshold ($\sigma_\beta^2 \approx 4$) results in an increase in the uncertainty of estimates. Beyond this threshold, as the prior becomes less informative, convergence issues arise due to nonidentifiability. It is worth noting that if the variance is too small, essentially reducing the prior toward a hard constraint also results in sampling issues such as poor mixing.

The Poisson model with the explicit parametrization of propensity parameters, with the above prior specification, is compared with the simplified negative binomial model (referred to as the overdispersed model) in 3 with the same priors on α_i's and β_k's and uniform priors over $\frac{1}{\omega_k}$. Figure 1 presents the plots of expected vs observed outcomes under the 2 models. The explicit Poisson model clearly provides a better fit to the data. We speculate that the Poisson model provides a better fit by capturing the variability in the data through the propensity parameters. The negative binomial model, on the other hand, is a reparametrization obtained by integrating the propensity parameters with respect to a prior whose variance is defined as a function of other parameters. Therefore, the negative binomial model can be interpreted as a constrained version of the original model, which fails to capture the variability in the data as well.

The degree distribution estimated from the 2 models is presented in Figure 2A. The degree distributions are very close

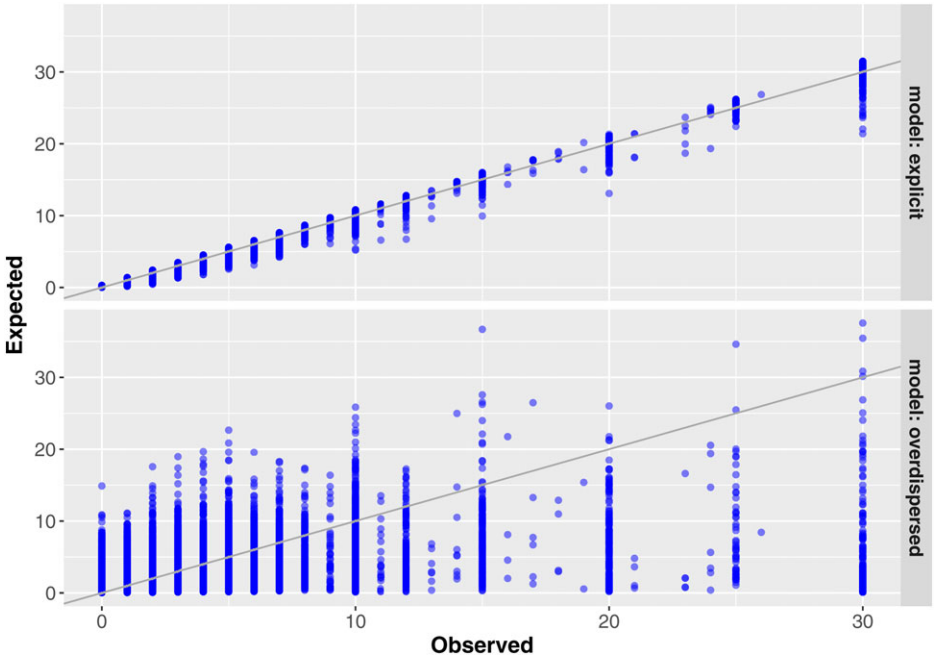

FIGURE 1 The expected vs observed outcomes under the explicit Poisson model (top panel) and the overdispersed negative binomial model (bottom panel)

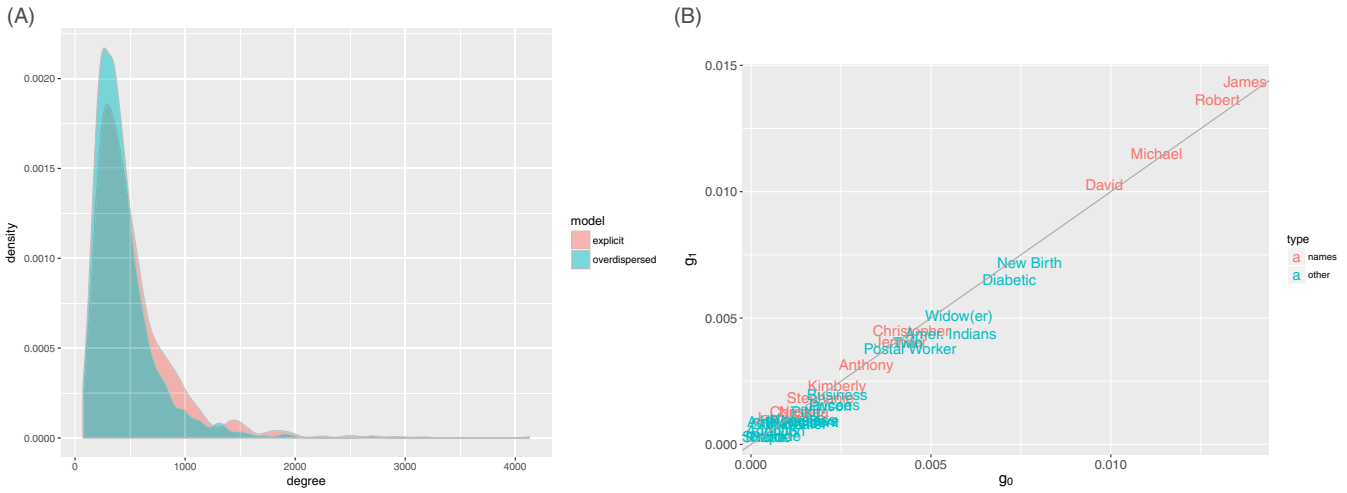

FIGURE 2 Comparison of (A) degree distributions and (B) prevalence estimates obtained by the overdispersed model and the explicit parametrization of individual propensities—the groups in red are the auxiliary subpopulations

under the 2 models, with a slightly higher variation under the explicit model due to more flexibility. In Figure 2B), the prevalence estimates are compared for the 32 subpopulations under the 2 models where the estimates are in agreement, except for the estimation error.

The link between the 2 models is investigated by studying the relationship between the unshared parameters, that is, the overdispersion and the propensity parameters. Figure 3 shows the overdispersion parameter estimates plotted against the variance of propensity estimates within the subpopulations. The positive correlation between the propensity variance estimates and the overdispersion parameters is interpreted as follows: Overdispersion represents the variation among individuals in their number of acquaintances in a certain subpopulation. This is directly related to the variation in individual propensities to make ties with a group when adjusted for the degrees and prevalence of the group. Therefore, the variance of propensity scores is an equivalent measure to overdispersion.

Note that the main goal of this section is to demonstrate the role of informative prior distributions in resolving nonidentifiability and using more interpretable model parametrization. Introducing external information through the prior is a generalizable approach and can be easily implemented for different problems.

3 | A BAYESIAN MODEL FOR PARTICLE DETECTION

Consider a Bayesian analysis of the data generated by particle detectors for the purpose of discovering the Higgs particle.

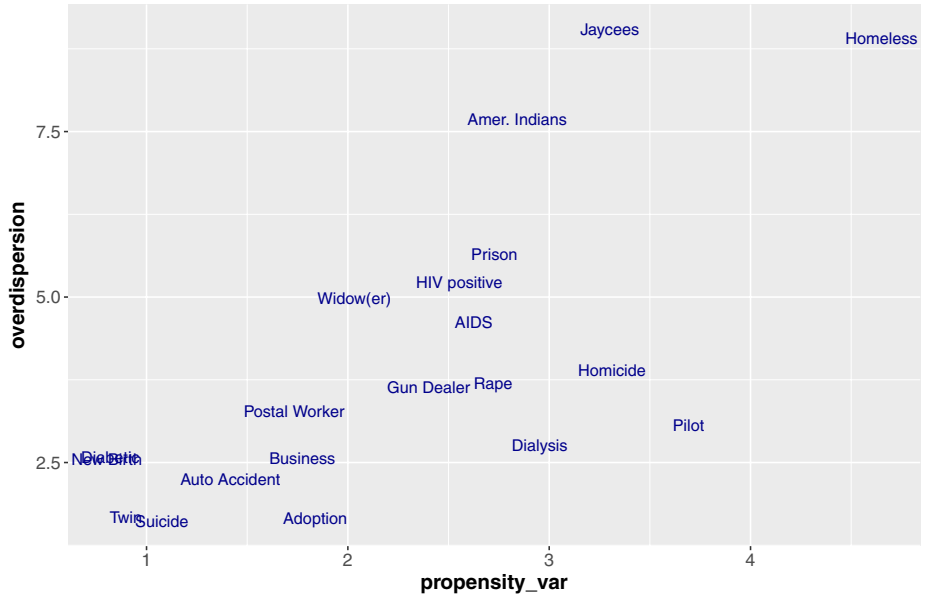

FIGURE 3 Overdispersion parameter estimates against the variance of the estimates of propensities for each subpopulation

While having significant scientific implications, the discovery of the Higgs boson is an interesting and complex statistical analysis problem. The key feature of the Higgs problem is that in addition to the experimental data there exist other sources of information that should be used for discovering new phenomenon. The theory and Monte Carlo studies provide information that is crucial for the detection of the Higgs particle. In the following, we provide a simplified description of the experiments, data, and the statistical analysis of the discovery of the Higgs particle. We then review a Bayesian hierarchical model proposed by Golchi and Lockhart [13] and explain that the use of informative priors in a Bayesian framework can play an important role in most effectively combining information from multiple sources to separate signal from noise.

The Standard Model (SM) of particle physics describes the dynamics of subatomic particles. The Higgs particle is an essential component of the SM [8,1,7,2]. 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, a high-energy collider that is specifically designed and constructed to detect the Higgs particle.

A simplified description of the experiment is as follows [13]. Beams of protons circulate at very high speeds in the LHC and collide inside two detectors (ATLAS and CMS). Collisions or "events" result in the generation of new particles that possibly include the Higgs boson. The Higgs particle, if generated, decays extremely quickly into other known SM particles and, therefore, cannot be detected directly. The existence of the Higgs particle is inferred by the combinations of detectable particles predicted by the SM. Once a Higgs particle is created, there are various "decay modes" through which it may decay. The decay process is reconstructed based on the detected collision byproducts. Events whose reconstructed processes match one of the possible Higgs decay modes are labeled "Higgs candidates". The mass of the unobserved particle is then computed from the reconstruction, and a histogram of the estimator of mass is created. There are other processes, not involving the Higgs boson, that can result in Higgs events; these are referred to as "background events". Therefore, the histogram created is either a histogram of background-only events if the Higgs particle does not exist or a histogram of background-plus-signal events otherwise.

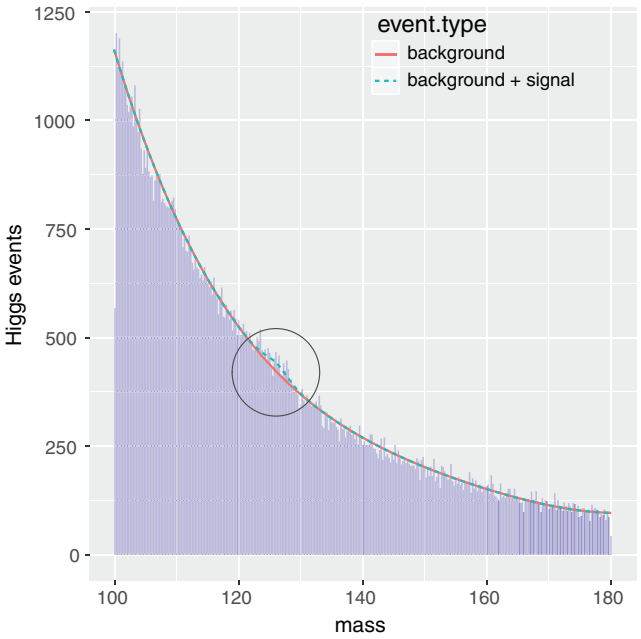

FIGURE 4 Simulated data representing the corresponding potential Higgs mass for each event. The circle identifies the overflow of events that is considered evidence of the existence of the Higgs particle

Figure 4 shows typical background-plus-signal data that are generated by computer models that simulate the behavior of particle detectors. The slight overflow of events around $m = 126$ (within the circle) shows the level of evidence that

generally appears in such data. A glance at these data suggests that, without any additional information about the distribution of the background events or the shape and size of the overflow (signal function), little statistical power is available in signal detection. In fact, the strong evidence reported in physics papers of the existence of the Higgs particle (the so-called 5-σ effect) is not obtained from one set of experimental data alone—the discovery of the Higgs boson was possible by combining data from various experiments together with Monte Carlo data, information given by the theory (SM), and other model assumptions.

In the current procedure, a parametric form is assumed for the background function whose parameters are estimated from a combination of real data and Monte Carlo samples generated from LHC simulators [2]. Different parametric background models are used for different decay channels to capture the specific background-event distribution in that channel. The estimated background model is then used in the analysis.

Another important component of the analysis of Higgs data is the information provided by theory about the shape and size of the signal for a given mass of the Higgs particle. For any mass value, the SM predicts the distribution of signal events in the form of a fully specified function [8,1]. However, in the current procedure, the search is not restricted to the SM prediction: a normalizing factor, referred to as "signal strength," is used as a free parameter, thereby extending the search for signals of different magnitudes than that of the SM-predicted signal.

Therefore, the background and signal are both partially specified but are controlled by free parameters. The important question is how to use the information available about the background and signal with an adequate level of uncertainty. Incorrect or insufficient information about the background can easily result in the failure of signal detection. A misspecified background distribution increases the probability of false discovery. On the other hand, if the SM prediction is used strictly (fixed signal strength parameter), and little flexibility is allowed for the background, the detection of a signal that is slightly different than that predicted by the theory, due to experimental or observational errors, can be extremely difficult.

In a Bayesian framework, the problem translates into prior specifications for the background as well as the signal strength parameter. In the following, we review the model given by Golchi and Lockhart [13] together with the prior choices and discuss how alternative prior choices can affect the results of the analysis. This brief sensitivity analysis is meant to emphasize that informative priors, if properly specified to correctly reflect the available information, are formal and statistically justifiable tools for incorporating external knowledge to a statistical model with an adequate level of uncertainty.

In Golchi and Lockhart [13], the data (presented in Figure 4) are modeled as realizations of a Poisson process whose intensity function is given by the sum of a background process $\Lambda(m)$ and a signal function $s_H = \mu s_{m_H}(m)$, where μ is the signal strength parameter. The shape of the signal function is given by theory, and its location is determined by the parameter of interest, the unknown mass of the Higgs particle $m_H \in M$, where $M = \{\emptyset\} \cup (m_0, m_n)$ $((m_0, m_n) \subset \mathcal{R}^+ - \{0\})$. Having $m_H \in (m_0, m_n)$ 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). The search window implied by the available data is (100, 180). The uncertainty about background, $\Lambda(m)$, is modeled by a log-Gaussian process:

$$\log \Lambda_{\eta,\sigma^2}(m) \sim GP(\xi(m), \rho_{\eta,\sigma_2}(m, m')), \qquad m \in (m_0, m_n), \tag{8}$$

where the mean function is specified as

$$\xi(m) = \log(g(m)) - \frac{\sigma^2}{2}. \tag{9}$$

This parametrization is used to take advantage of the typical parametric forms currently used to model the background function [1] for the expectation of background, $E(\Lambda)$. Examples of such parametric forms are Bernstein polynomials and exponential decay functions. The notation, $g(m)$, therefore, represents a model assumed for Λ, which is transformed through 9 to specify the mean of $\log(\Lambda)$.

The covariance function is given by:

$$\rho_{\eta,\sigma_2}(m, m') = \sigma^2 \exp(-\eta(m - m')^2), \tag{10}$$

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

The signal function used is a Gaussian probability density function with the location parameter m_H:

$$s_{m_H}(m) = c_{m_H} \phi\left(\frac{m - m_H}{\varepsilon}\right), \quad m_H \in (m_0, m_n); \qquad s_\emptyset(m) = 0, \tag{11}$$

where c_{m_H} is a scaling constant, and ϕ is the normal probability density function with standard deviation ε. Note that in the current practice [1,8], a slightly more complex signal shape called the "crystal ball function" is used. However, we did not have access to the required information to recover the actual signal function. The parameters of the above signal function were learned using histograms of signals at 3 selected mass values. For a more detailed description of signal construction, see Golchi and Lockhart [13].

The likelihood is given by:

$$\pi(\mathbf{y} \mid \Lambda, m_H) = \prod_{i=1}^{n} \frac{\exp(-\Gamma_i)\Gamma_i^{y_i}}{y_i!}, \tag{12}$$

where

$$\Gamma_i = \int_{m_{i-1}}^{m_i} [\Lambda(m) + \mu s_{m_H}(m)] dm. \tag{13}$$

The grid $\mathbf{m} = (m_0, m_1, \ldots, m_n)$ is the vector of bin boundaries over the search window.

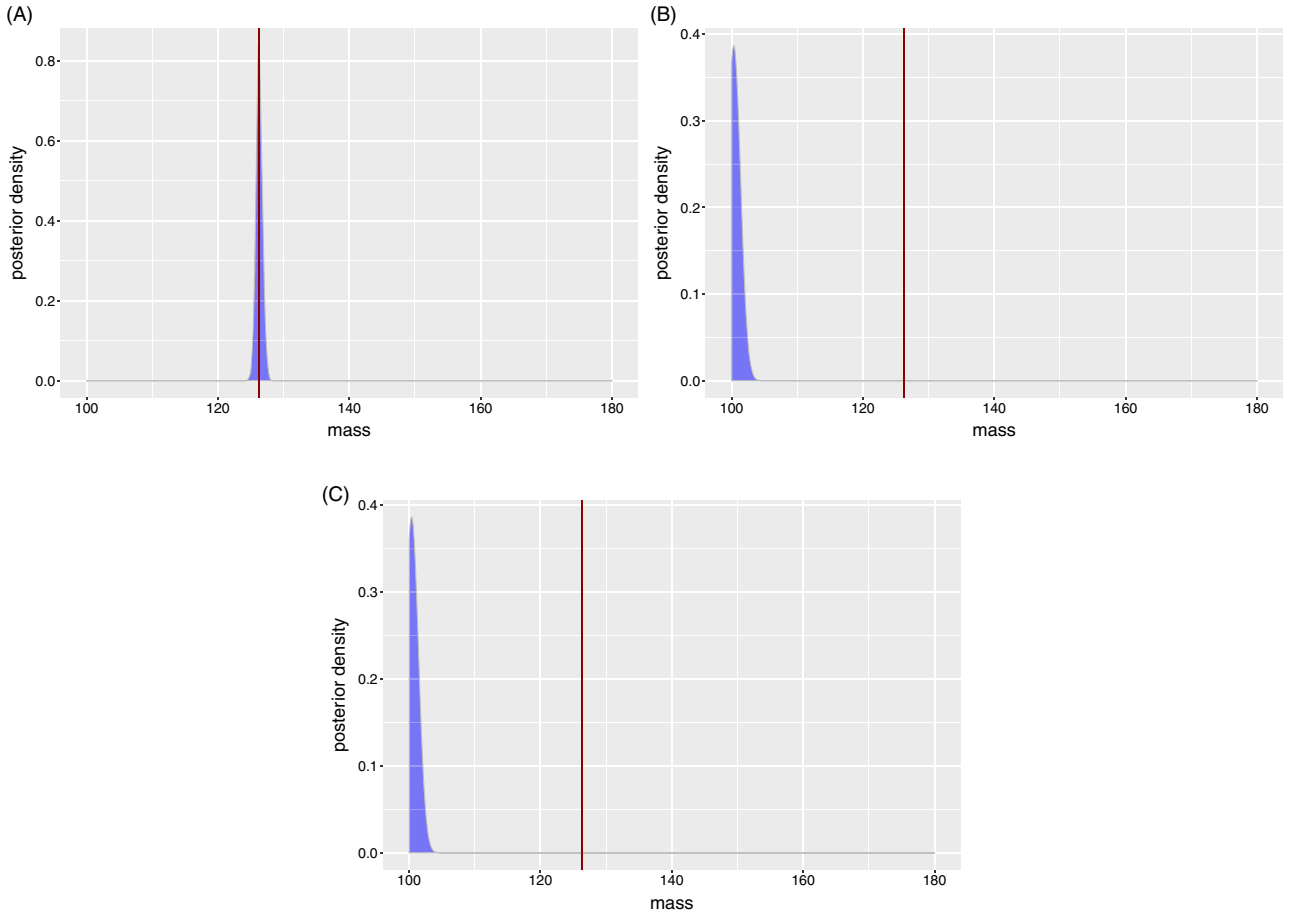

FIGURE 5 Kernel density estimates of the posterior distribution of the mass of the Higgs particle with (A) informative prior, (B) somewhat informative prior, and (C) noninformative prior on the background. The vertical red lines are drawn at the reported mass of the Higgs particle ($m \approx 126$)

The posterior distribution of the model parameters $\theta = (\alpha, \beta, \eta, \sigma^2, \Lambda, m_H, \mu)$, given the data **y,** is as follows:

$$\pi(\theta \mid \mathbf{y}) = \frac{\pi(\theta)\pi(\mathbf{y} \mid \theta)}{\int \pi(\theta)\pi(\mathbf{y} \mid \theta) d\theta}. \tag{14}$$

The prior is:

$$\pi(\theta) = \pi(\alpha)\pi(\beta)\pi(\eta)\pi(\sigma^2)\pi(\Lambda \mid \alpha, \beta, \eta, \sigma^2)\pi(m_H)\pi(\mu). \tag{15}$$

The prior distribution $\pi(m_H)$ is a mixture of a point mass at $m_H = \emptyset$ and a continuous distribution on (m_0, m_n). The hyperparameters β_i are assigned normal priors, and the hyperpriors on η and σ^2 are inverse Gamma distributions with shape and scale parameters equal to 1.

In Golchi and Lockhart [13], a fourth-order Bernstein polynomial is used to specify the background:

$$g_0(m) = \sum_{i=0}^{4} \beta_i h_i(z(m)), \tag{16}$$

where $z : (m_0, m_n) \to (0, 1)$ is an affine transformation of mass onto the unit interval, and the basis functions are given by:

$$h_i(z) = \binom{4}{i} z^i (1 - z)^{4-i} \qquad z \in (0, 1). \tag{17}$$

Here, to emphasize the importance of prior specification, we use 2 different models in addition to the 1 used in Golchi and Lockhart [13]. First, we use an exponential decay function to define the background mean:

$$g_1(m) = \alpha \exp(-\beta z(m)). \tag{18}$$

Second, we use a constant mean function:

$$g_2(m) = \alpha. \tag{19}$$

The above alternative mean functions may be considered misspecified informative priors.

The sequential Monte Carlo (SMC) algorithm outlined in Algorithm 1 in Golchi and Lockhart [13] is used to sample from the posterior distribution 14 with the 3 background mean choices mentioned above. Figure 5 shows kernel density estimates of the marginal posterior distribution of the mass over the search window under each of the discussed prior specifications. The posterior distribution is focused around the reported mass of the Higgs particle under the background prior with the Bernstein polynomial mean function (Figure 5A). The remaining search window, (m_0, m_n), is excluded as the possible mass of the discovered particle.

Under the background prior with the exponential mean function (Figure 5B) and the constant mean function, however, the posterior is entirely mislead toward the fluctuations in the data at the left edge of the search window. As the model is misinformed about the assumed shape of the background,

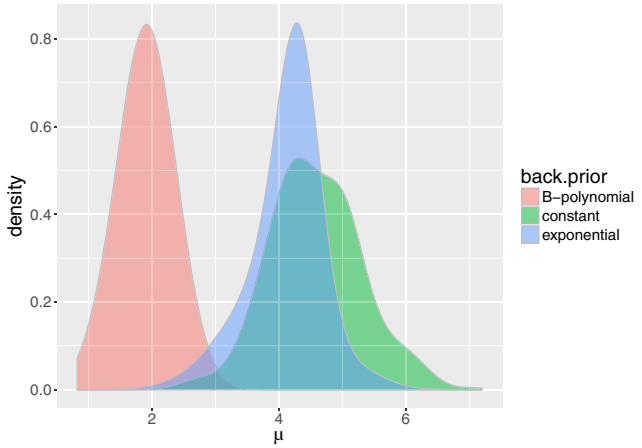

FIGURE 6 Kernel density estimates of the marginal posterior distribution of signal strength under 3 different parametrizations of the background mean

the signal is completely missed as a feature of the background curve. The credible interval for the mass misses the reported mass of the Higgs boson, showing the sensitivity of the results to assumptions made about the background in this problem.

Note that the prior distribution over the signal strength parameter μ can play an important role as well. However, there is an emphasis on treating this parameter as a "free" parameter. Therefore, a rather defuse prior on μ is chosen, and the focus of this discussion is on the background prior. It is, however, interesting to observe the interplay of mass and signal strength posterior distributions as a result of each background mean formulation. Figure 6 shows kernel density estimates of the marginal posterior distribution of signal strength under the three background mean parametrizations.

4 | INFERENCE FOR ORDINARY DIFFERENTIAL EQUATIONS

Consider the problem of estimating the parameters of a system of ordinary differential equations (ODE) from noisy observations. In the most real problems, the ODE cannot be solved analytically, and evaluation of the likelihood relies on numerical solutions that, in turn, rely on the ODE parameters and initial states. Inference in such scenarios is challenging due to the sensitivity of the likelihood to small changes in the parameters. In Bayesian inference, these challenges are translated into highly peaked and often multimodal posterior surfaces that are very difficult to explore using MCMC [4,6].

In Golchi and Campbell [12], Bayesian inference for ODEs is addressed by relaxation of adherence to the ODE. The data are modeled as the summation of the ODE solution and a discrepancy term that facilitates computation by fitting the deviation of data from the ODE solution for "incorrect" parameter values. Using sequential Monte Carlo, Golchi and Campbell [12] propose reducing the role of the discrepancy at each step until it completely diminishes from the model, and full adherence to the ODE is achieved.

The discrepancy term used by Golchi and Campbell [12] is a Nadaraya–Watson kernel smoother whose contribution to the model is controlled by the bandwidth and a coefficient parameter. The introduction of a kernel smoother as the discrepancy term is merely a technique to create a specialized implementation of SMC in the present framework. While Golchi and Campbell [12] provide an intuitive explanation for the performance of the proposed approach, no Bayesian interpretation can be made about this technique.

In this section, we propose replacing the kernel smoother with a zero-mean Gaussian process (GP) that essentially performs similarly in the relaxation of adherence to the ODE but is also interpretable as a probability distribution around the ODE solution. In other words, Bayesian computation is facilitated by departing from the model adherence assumption via a prior distribution that is made more informative sequentially at each step, converging to a point mass at the ODE solution.

The example used in Golchi and Campbell [12] is the analysis of data from the second outbreak of the black plague from June 19, 1666 to November 1, 1666 in the village of Eyam, UK. The data include the cumulative number of deaths at specific time points. The villagers voluntarily quarantined themselves to prevent the spread of the disease to neighborhood villages, and therefore, the population size is considered finite during the time window covered by the data. An epidemiological model known as a Susceptible-Infected-Removed (SIR) is assumed to underlie the data. Under the SIR model, at a given time, ν, the population is split into groups of Susceptible $S(\nu)$, Infected, $I(\nu)$, and Removed, $R(\nu)$ [5,19]. As there is no recovery from the plague, $R(\nu)$ is the number of deaths up to time ν. The rates of change of states $S(\nu)$, $I(\nu)$, and $R(\nu)$ are given by the following system of ODEs:

$$\begin{aligned} \frac{dS(\nu)}{d\nu} &= -\beta S(\nu)I(\nu), \\ \frac{dI(\nu)}{d\nu} &= \beta S(\nu)I(\nu) - \alpha I(\nu), \\ \frac{dR(\nu)}{d\nu} &= \alpha I(\nu), \end{aligned} \tag{20}$$

where β describes the plague transmissivity, and α represents the rate of death once an individual is infected by plague.

At time 0, the population only consists of susceptible and infected individuals. Therefore, $R(0) = 0$ and $S(0) = N - I(0)$. Consequently, $I_0 = I(0)$ is included in the vector of unknown parameters to be estimated: $\theta = (\alpha, \beta, I_0)$.

With a finite population of size N, the n observed cumulative deaths $y(\nu)$ at times $\{\nu_1, \ldots, \nu_n\}$ are modeled as binomial counts whose expected value is the solution to $R_\theta(\nu)$ from 20:

$$P(\mathbf{y} \mid R_\theta(\nu)) = \prod_{i=1}^{n} \binom{N}{y_i} \left(\frac{R_\theta(\nu_i)}{N}\right)^{y_i} \times \left(1 - \frac{R_\theta(\nu_i)}{N}\right)^{(N-y_i)}.$$

Following the approach proposed by Golchi and Campbell [12], as a relaxation of the ODE adherence assumption, $R_\theta(\nu)$ in the likelihood is replaced by $R_\theta(\nu)$ plus a discrepancy term. However, a zero-mean GP is used here instead of the kernel smoother to make the approach interpretable

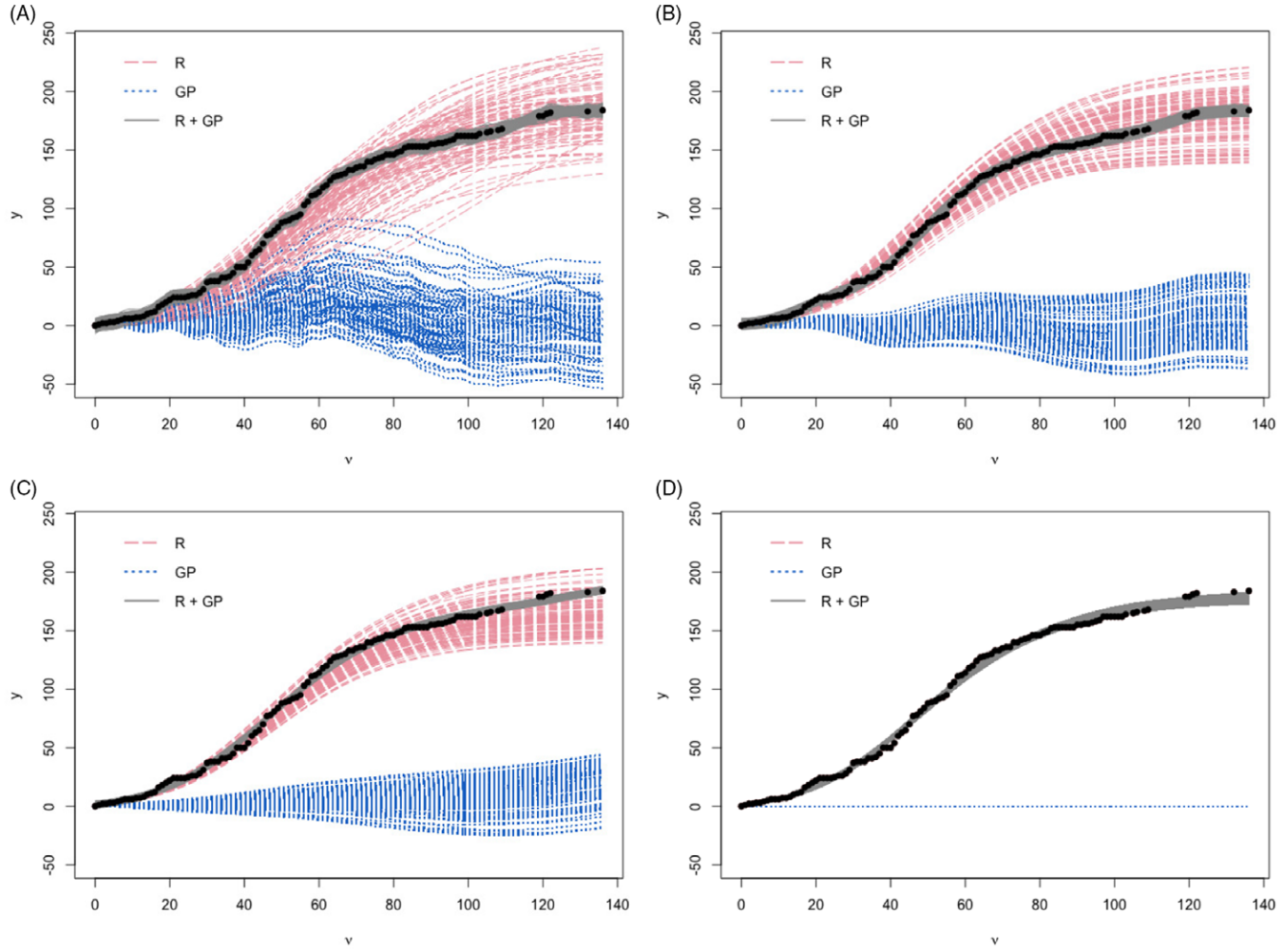

FIGURE 7 Posterior sample paths (solid gray lines) generated as the sum of the ordinary differential equations solution (dashed red lines) and a zero-mean Gaussian process (GP) (dotted blue lines) at (A) step 1, (B) step 7, (C) step 15, and (D) step 25 (final step) of the sequential Monte Carlo sampling. The black dots are the observations

in a Bayesian framework. Every evaluation of the likelihood requires numerically solving 20 to obtain $R(\theta, \nu)$ and fitting a zero mean GP to the residuals, $y(\nu) - R(\theta, \nu)$. More explicitly, for each set of parameter values θ, a sample path is given by

$$\widehat{R}_{\theta,\xi}(\nu) = R_\theta(\nu) + z_\xi(\nu), \tag{21}$$

where $z_\xi(\nu)$ is a zero-mean GP with a covariance function parametrized by $\xi = (\eta, \sigma^2, \tau^2)$,

$$cov(z_\xi(\nu), z_\xi(\nu')) = \begin{cases} \sigma^2 \exp(-\eta(\nu - \nu')^2) & \text{if } \nu \neq \nu', \\ \sigma^2 + \tau^2 & \text{if } \nu = \nu'. \end{cases} \tag{22}$$

The GP is interpreted as a prior distribution over the ODE solution that can be tuned to allow departures from the ODE adherence assumption. SMC is used to sample from the target posterior distribution $\pi_T(\boldsymbol{\theta}|\mathbf{y})$, where the intermediate densities are defined by a sequence of covariance parameter values of the GP:

$$\pi_t(\theta, \widehat{R} \mid \theta) \propto P(\mathbf{y}|\widehat{R}_\theta)\pi(\widehat{R}_{\theta,\xi_t} \mid R_\theta)\pi(\theta), \tag{23}$$

where $\pi(\widehat{R}_{\theta,\xi_t} \mid R_\theta)$ represents the GP prior, and $\pi(\boldsymbol{\theta})$ is the product of independent priors on α, β, and I_0. As t increases, the parameters ξ_t are adjusted to make the prior more focused about the ODE solution by making the GP sample paths smoother (decreasing η_t) and the variance smaller (decreasing σ_2). Eventually, the GP prior converges to a point mass at the ODE solution, that is, $\pi(\widehat{R}_{\theta,\xi_t} \mid R_\theta) = 1$.

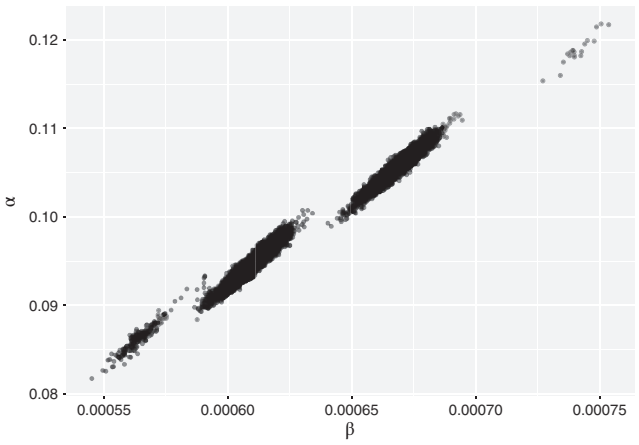

FIGURE 8 Samples from the joint posterior distribution of the model parameters. The 3 large clouds of particles correspond to $I_0 = 6$, $I_0 = 5$, and $I_0 = 4$, respectively, from left to right

Figure 7 shows 4 selected steps along the 25 steps of the SMC sampler. The evolution of the sample paths (solid gray lines) that are obtained by correcting the ODE solution (dashed red lines) by a GP fit to the deviations of the observations from the ODE solution (dotted blue lines) is illustrated in these sample plots. At the first step (Figure 7A), the GP interpolates the residuals after subtracting the ODE solution from the observations, thereby preventing the likelihood from crashing to zero and allowing exploration of the parameter space. The role of the GP is weakened throughout the sampling by switching from interpolation to smoothing (Figure 7B and C). Finally, the GP is eliminated from the likelihood, and the final sample paths and the ODE solutions become the same (Figure 7D).

Figure 8 shows samples from the final joint posterior of α and β in the form of separate clouds that correspond to initial values I_0. This figure closely resembles the results of Golchi and Campbell [12].

5 | CONCLUSION

In this article, the role of informative priors in Bayesian inference and computation is emphasized through 3 different example scenarios. Each example represents a challenging statistical problem. The problems are addressed in a Bayesian framework, and the contribution of prior distributions in overcoming the challenges in inference and computation is discussed.

In the first example, Bayesian analysis of aggregated relational data is considered. The main challenge is that the proposed model is nonidentifiable, and inference cannot be carried out without taking advantage of additional information besides the data. Therefore, the choice of an informative priors plays a crucial role from both inference and computational point of views—external information that is required to identify the model are incorporated in a formal and interpretable manner resulting in a hierarchical model that can be implemented in Stan and made accessible to users. We note that the same strategy may be used to gain computational efficiency and improvements in the inference for related models in Mccormick and Zheng [16], McCormick et al. [18], and McCormick and Zheng [17].

The second example is a statistical analysis for the discovery of the Higgs particle in a Bayesian framework. This example represents an interesting scenario where incorporating information from multiple sources besides the data, such as the theory and other experiments, has a crucial effect on the answer to an important scientific question. Discovery of new phenomenon would not have been possible without adequate usage of all the available information with appropriate level of uncertainty. This is demonstrated by comparing the results of the analysis with 3 prior specifications for the nuisance parameters.

In the third example, we use an informative prior from a purely computational perspective. As mentioned before, unlike the first 2 examples, the prior is not a part of the model and is used merely as a computational tool. The problem is to make an inference about the parameters of a system of ordinary differential equations where the assumption of adherence to the ODE results in difficulties in MCMC sampling from the posterior distribution of ODE parameters. A prior distribution is used to allow temporary departure from the ODE to facilitate exploration of the parameter space. In a sense, by considering the model assumption as a highly informative prior, relaxation from the assumptions is achieved by adding uncertainty through the prior distribution. However, the prior uncertainty is eventually eliminated from the model, thereby assuring full adherence to the underlying ODE.

ACKNOWLEDGMENT

This research is partially supported by NSF grant SES-1023176 and the Natural Sciences and Engineering Research Council of Canada (NSERC). The author thanks Dr Andrew Gelman and Dr Tian Zheng for their support and invaluable comments.

ORCID

Shirin Golchi http://orcid.org/0000-0003-3382-9563

REFERENCES

1. ATLAS Collaboration, *Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B 716(1) (2012), 1–29.
2. ATLAS Collaboration, *Measurement of higgs boson production in the diphoton decay channel in* pp *collisions at center-of-mass energies of 7 and 8 tev with the atlas detector*, Phys. Rev. D 90 (2014), 112015.
3. M. J. Bayarri and J. O. Berger, *The interplay of Bayesian and frequentist analysis*, Stat. Sci. 19 (2004), 58–80. MR2082147
4. B. Calderhead and M. Girolami, *Estimating bayes factors via thermodynamic integration and population mcmc*, Comput. Statist. Data Anal. 53 (2009), 4028–4045. MR2744303
5. D. A. Campbell and S. Lela, *An anova test for parameter estimability using data cloning with application to statistical inference for dynamic systems*, Comput. Statist. Data Anal. 70 (2014), 257–267. MR3125492
6. D. A. Campbell and R. J. Steele, *Smooth functional tempering for nonlinear differential equation models*, Stat. Comput. 22 (2011), 429–443. MR2865027
7. CDF Collaboration, *Search for high mass resonances decaying to muon pairs in* $\sqrt{s} = 1.96$ TeV$p\bar{p}$ *collisions*, Phys. Rev. Lett. 106 (2011), 121801.
8. CMS Collaboration, *Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc*, Phys. Lett. B 716(1) (2012), 30–61.
9. A. Gelman, *Bayesian model-building by pure thought: Some principles and examples*, Statist. Sin. 6 (1996), 215–232. MR1379058
10. A. Gelman, *Objections to Bayesian statistics*, Bayesian Anal. 3 (2008), 445–450. MR2434394
11. A. Gelman et al., *A weakly informative default prior distribution for logistic and other regression models*, Ann. Appl. Stat. 2 (2008), 1360–1383. MR2655663
12. S. Golchi and D. Campbell, *Sequentially constrained Monte Carlo*, Comput. Statist. Data Anal. 97 (2016), 98–113. MR3447039

13. S. Golchi and R. Lockhart, A frequency-calibrated Bayesian search for new particles, Annals of Applied Statistics. In Press. arXiv:1501.02226 [stat.AP], 2016.
14. R. J. Little, *Calibrated bayes: A bayes/frequentist roadmap*, Amer. Statist. 60 (2006), 213–223. MR2246754
15. C. McCarty et al., *Comparing two methods for estimating network size*, Hum. Organ. 60 (2001), 28–39.
16. T. H. McCormick and T. Zheng, *Adjusting for recall bias in 'how many xs do you know?' surveys*. Proc. Joint Statistical Meetings, 2007.
17. T. H. McCormick and T. Zheng, *Latent demographic profile estimation in hard-to-reach groups*, Ann. Appl. Stat. 6 (2012), 1795–1813. MR3058684
18. T. H. McCormick, M. J. Salganik, and T. Zheng, *How many people do you know?: Efficiently estimating personal network size*, J. Amer. Statist. Assoc. 105 (2010), 59–70. MR2757192
19. G. F. Raggett, *Modelling the eyam plague*, Inst. Math. Appl. 18 (1982), 221–226.
20. P. B. Stark, *Constraints versus priors*, SIAM/ASA J. Uncertain. Quantif. 3 (2015), 586–598. MR3372107
21. T. Zheng, M. J. Salganik, and A. Gelman, *How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks*, J. Amer. Statist. Assoc. 101 (2006), 409–423. MR2256163

How to cite this article: Golchi S. *Informative priors in Bayesian inference and computation*, *Statistical Analysis and Data Mining: The ASA Data Science Journal* 2019;12:45–55. https://doi.org/10.1002/sam.11371.