Choosing priors in Bayesian ecological models by simulating from the prior predictive distribution

Jeff S. Wesner and Justin P.F. Pomeranz

University of South Dakota, Department of Biology, Vermillion, SD 57069

Jeff.Wesner@usd.edu

Abstract

Bayesian data analysis is increasingly used in ecology, but prior specification remains focused on choosing non-informative priors (e.g., flat or vague priors). One barrier to choosing more informative priors is that priors must be specified on model parameters (e.g., intercepts, slopes, sigmas), but prior knowledge often exists on the level of the response variable. This is particularly true for common models in ecology, like generalized linear mixed models that have a link function and potentially dozens of parameters, each of which needs a prior distribution. We suggest that this difficulty can be overcome by simulating from the prior predictive distribution and visualizing the results on the scale of the response variable. In doing so, some common choices for non-informative priors on parameters can easily be seen to produce biologically impossible values of response variables. Such implications of prior choices are difficult to foresee without visualization. We demonstrate a workflow for prior selection using simulation and visualization with two ecological examples (predator-prey body sizes and spider responses to food competition). This approach is not new, but its adoption by ecologists will help to better incorporate prior information in ecological models, thereby maximizing one of the benefits of Bayesian data analysis.

Keywords: *Bayesian, prior predictive distribution, GLMM, simulation*

Introduction

The distinguishing feature between Bayesian and non-Bayesian statistics is that Bayesian statistics treats unknown parameters as random variables governed by a probability distribution, while non-Bayesian statistics treats unknown parameters as fixed and unknown quantities (Ellison 2004, Hobbs and Hooten 2015). A common misconception is that only Bayesian statistics incorporates prior information. However, non-Bayesian methods can and often do incorporate prior information, either informally in the choices of likelihoods and model structures, or formally as penalized likelihood or hierarchical modeling (Hobbs and Hooten 2015, Morris et al. 2015).

While prior information is not unique to Bayesian models, it is required of them. For example, in a simple linear regression of the form $y \sim \text{Normal}(\alpha + \beta x, \sigma)$, the intercept α, slope β, and standard deviation σ are unknown parameters that each need a prior probability distribution. There are differing opinions and philosophies on the best practices for choosing priors (Lindley 1961, Edwards et al. 1963, Morris et al. 2015, Wolf et al. 2017, Gelman et al. 2017, Lemoine 2019, Banner et al. 2020). In ecology, a common practice is to assign so-called non-informative priors that effectively assign equal probability to all possible values using either uniform or diffuse normal priors with large variances (Lemoine 2019). These priors allow Bayesian inference to proceed (i.e., produce a posterior distribution), but with presumably limited influence of the priors (Lemoine 2019).

Reasons for using non-informative priors are varied but are at least in part driven by a desire to avoid the appearance of subjectivity and/or a reliance on default settings in popular software (Gelman and Hennig 2017, Banner et al. 2020). There are several arguments against this approach. First, "non-informative" is a misnomer. All proper priors influence the posterior distribution to some extent (Hobbs and Hooten 2015). As a result, a prior cannot just be assumed as non-informative based on default settings or a wide variance (Seaman III et al. 2012). Its implications for the model should be checked just like any other subjective assumption in data analysis, whether Bayesian or not (Gelman et al. 2017, Banner et al. 2020). Second, adhering to non-informative priors removes

a major potential benefit of Bayesian analysis, which is to explicitly incorporate prior research and expertise into new science (Hobbs and Hooten 2015, Lemoine 2019, Rodhouse et al. 2019). Third, informative priors can help to reduce spurious conclusions due to errors in magnitude or sign of a relationship by treating extreme values in the data skeptically (Gelman et al. 2012, Lemoine 2019). Finally, informative priors make computational algorithms like MCMC run more efficiently, which can save hours or days of computing time in complex models (Hobbs and Hooten 2015). An additional way to improve efficiency can come from different choices of prior distributions, such as an inverse-gamma distribution on the variance rather than the exponential prior on the standard deviation that we fit in this model. For more complete discussion on this, see Gelman and others (2006).

While there are clear arguments for why ecologists *should* use more informative priors, it is often difficult to know *how* to use them. Even for seemingly simple and routine models, like logistic or Poisson regression, it can be difficult to understand *a priori* how priors affect the model, because they must be assigned in the context of likelihood with a linearizing link-function (Seaman III et al. 2012, Gelman et al. 2017). In other words, prior specification takes place on model parameters (e.g., slopes, intercepts, variances), but prior knowledge is often easier to assess on the model outcomes (Kadane et al. 1980, Bedrick et al. 1996, Gabry et al. 2019). This is particularly true for models that are commonly used in ecology, such as generalized linear mixed models with interactions. These models may have dozens of parameters and hyperparameters, each of which require a prior probability distribution (Bedrick et al. 1996, McElreath 2020).

We suggest that ecologists can address this problem by simulating from the prior predictive distribution and visualizing the implications of the priors on outcomes of interest (e.g., means and confidence intervals of treatment groups, simulated data, or regression lines). In this paper, we demonstrate this approach using two case studies with ecological data (Figure 1). All data and code are available at: https://github.com/jswesner/prior_predictive.

Prior Predictive Simulation

An attractive feature of the Bayesian approach is that the models are generative. This means that we can simulate potential data from the model so long as the parameters are assigned a proper probability distribution (Gelman et al. 2013). This feature is routinely used to check models and prior influence *after* fitting the data using the posterior predictive distribution (Lemoine 2019, Gelman et al. 2020), but it can also be used before seeing the data using the prior predictive distribution (Gabry et al. 2019).

The general workflow for prior predictive simulation is:

1) Draw N realizations from a prior distribution
2) For each draw, simulate a model outcome or new data from the likelihood
3) Plot the results
4) Use domain knowledge to assess whether simulated values reflect prior knowledge
5) If simulated values do not reflect prior knowledge, change the prior distribution, likelihood, or both and repeat the simulation from step 1
6) If simulated values reflect prior knowledge, add the data and estimate the posterior distribution

This amounts to a prior predictive check to satisfy the expectation that "simulations from the full Bayesian model…should be plausible data sets" (Kennedy et al. 2019). We demonstrate it with two motivating examples.

Example 1: Predator-Prey Body Sizes - Simple Linear Regression

Data - Understanding predator-prey interactions has long been a research interest of ecologists. Body size is related to a number of aspects that influence these interactions. For example, predators are often gape-limited, meaning that larger predators should be able to eat larger prey. The data set

of Brose et al. (2006) documents over 10,000 predator-prey interactions, including the mean mass of each (Figure 1a).

Model - We examined the hypothesis that the prey body mass increases log-linearly with predator body mass using a simple linear model:

$$\log(y_i) \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$\mu_i = \alpha + \beta \log(x_i) \tag{2}$$

$$\alpha \sim \text{Normal}(0, \sigma_\alpha) \tag{3}$$

$$\beta \sim \text{Normal}(0, \sigma_\beta) \tag{4}$$

$$\sigma \sim \text{Exponential}(\phi) \tag{5}$$

where $\log(y_i)$ is natural log transformed prey mass and $\log(x_i)$ is natural log transformed predator mass.

Priors - For the α and β priors, we first assign a mean of 0 with a "non-informative" standard deviation of 1000 [$N(0, 1000)$] (Table 1). These prior values are often used as defaults, especially in earlier Bayesian software to generate "flat" prior distributions and is commonly used in the ecological literature (McCarthy and Masters 2005, Banner et al. 2020). The mean of 0 in a normal distribution implies that the intercept and slope have equal probability of being positive or negative. For the exponential distribution, we specify an initial ϕ of 0.00001, chosen by plotting 100 simulations from the exponential function in R (R Core Team 2020) with varying values of ϕ [e.g., `plot(rexp(100, 0.00001)`]. A value of 0.00001 generated an average deviance of ~1,000 with values up to ~5,000, indicating the possibility of producing extremely large values.

After simulating regressions from these initial priors, we specified successfully tighter priors and repeated the simulations (Table 1; Figure 2). Those simulations were compared to reference points representing prior knowledge (Mass of earth, a Blue Whale, a virus, and a Carbon-12 atom). The

goal was to use these reference points to find a joint prior distribution that produced reasonable values of potential prey masses. We did this using two levels of the model (μ_i and y_i). For μ_i, we simulated 100 means across each value of x_i and plotted them as regression lines. For y_i, we simulated a fake data set containing simulated values of log prey mass for each of the 13,085 values of log predator mass (x_i) in the Brose et al. (2006) data.

Results - The weak "non-informative" priors make nonsense predictions (Figure 2a-c). In Figure 2a, all of the lines are impossibly steep, suggesting that predators could plausibly eat prey that are larger than earth or smaller than an atom. The stronger priors in Figure 2b suffer from the same problem, though the effect is less severe. The strongest priors (Figure 2c) produce more reasonable predictions, though they are still quite vague, with positive probability that predators could eat prey larger than an adult Blue Whale. The simulated fake data sets tell a similar story (Figure 2d-f), but with the added influence of σ (Equation 1).

We fit the model using the strongest prior set and overlaid these on the prior simulations (Figure 2c,f). As expected, there is a strong positive relationship between log predator and log prey size (Figure 2c - orange line), despite the uncertainty in the prior. The intercept is -4.8 ± 0.04 (mean ± sd), indicating that a predator weighing 1 gram (wet or dry mass) would eat prey 2-3 orders of magnitude smaller g. The slope is 0.6 ± 0.01, indicating a reliably positive relationship such that an increase in 1 log unit of predator mass corresponds with an increase in prey mass of 0.6 log units on average. Sigma is 3.7 ± 0.02, indicating an average residual for individual predator-prey data or +/- 3.7 log-units of prey mass. This is reflected in the simulated data, which show a wide range of simulated predator-prey size pairings, but all are within a reasonable range compared to prior predictions (Figure 2f).

There are several benefits to choosing a stronger prior. First, it is difficult to justify the two weakest priors on biological grounds. They place large amounts of prior probability on impossible values. This can matter when priors need to be justified to a granting agency or to reviewers. More critically, specification of priors can have conservation or legal implications, and the ability to justify priors

with simulation helps to improve transparency (Crome et al. 1996, Banner et al. 2020). Stronger priors also improve computational efficiency (McElreath 2020). We fit these models using the *brms* package (Burkner 2017). The algorithms associated with models that had the stronger or strongest priors were up to 50% faster than the model with weak priors, taking 56 vs 28 seconds on a standard laptop (compilation time + warmup time + sampling time). For more complex models with algorithms that take longer to run, this improvement can save hours or days of computing time.

Caveats - We know from the literature that predators are generally larger than their prey by 2-3 orders of magnitude (Trebilco et al. 2013). Therefore, it would make sense to alter the prior mean of the intercept to a value below zero, perhaps using an average predator/prey mass comparison from the literature. That is apparent from the prior versus posterior comparison in Figure 2c. Similarly, the fact that larger predators tend to eat larger prey is well-known, so the prior on the slope β could be changed to a positive mean. Another option is to standardize the predictor variable(s) ($x_{standardized} = \frac{x-\bar{x}}{\sigma_x}$) so that the regression slopes can be interpreted as units of standard deviation. This also improves interpretation of the intercept when values of zero (e.g., prey mass = 0) do not make sense (McElreath 2020). This would be most relevant in the current model if prey mass was included as raw mass, rather than log mass.

Example 2: Spider Abundance - Generalized Linear Mixed Model

Data - This data set comes from Warmbold and Wesner (2018), who measured terrestrial spider responses to different combinations of freshwater fish using fish enclosure cages in a backwater of the Missouri River, USA. They hypothesized that fish would reduce the emergence of adult aquatic insects by eating the larval stages in the water, causing a reduction in terrestrial spiders that feed on the adult forms of those insects. The original experiment contained six treatments. We present a simplified version comparing spider abundance above three treatments that contain either Smallmouth Buffalo (*Ictiobus bubalus*), Green Sunfish (*Lepomis cyanellus*), or a fishless control. Each treatment had four replicates for a total of 12 cages (each 2.3 m^2). The cages were arranged in