OIKOS

*Forum*

# Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses


**Nathan P. Lemoine**

*N. P. Lemoine (https://orcid.org/0000-0003-3436-7196) ✉ (lemoine.nathan@gmail.com), Dept of Biology, Marquette Univ., Milwaukee, WI, USA.*

**Oikos**
**128: 912–928, 2019**
doi: 10.1111/oik.05985

Subject Editor
and Editor-in-Chief: Dries Bonte
Accepted 21 March 2019

Throughout the last two decades, Bayesian statistical methods have proliferated throughout ecology and evolution. Numerous previous references established both philosophical and computational guidelines for implementing Bayesian methods. However, protocols for incorporating prior information, the defining characteristic of Bayesian philosophy, are nearly nonexistent in the ecological literature. Here, I hope to encourage the use of weakly informative priors in ecology and evolution by providing a 'consumer's guide' to weakly informative priors. The first section outlines three reasons why ecologists should abandon noninformative priors: 1) common flat priors are not always noninformative, 2) noninformative priors provide the same result as simpler frequentist methods, and 3) noninformative priors suffer from the same high type I and type M error rates as frequentist methods. The second section provides a guide for implementing informative priors, wherein I detail convenient 'reference' prior distributions for common statistical models (i.e. regression, ANOVA, hierarchical models). I then use simulations to visually demonstrate how informative priors influence posterior parameter estimates. With the guidelines provided here, I hope to encourage the use of weakly informative priors for Bayesian analyses in ecology. Ecologists can and should debate the appropriate form of prior information, but should consider weakly informative priors as the new 'default' prior for any Bayesian model.

Keywords: Bayesian statistics, frequentist statistics, Markov chain Monte Carlo, vague priors

Synthesis

Statistical practice in ecology is currently undergoing a paradigm shift as ecologists increasingly utilize Bayesian statistics to test their hypotheses. However, despite the explicit dependence of Bayesian statistics on priors, ecologists remain understandably wary of using informative priors and instead default to 'noninformative', 'vague', or 'flat' priors. Here, I demonstrate that using 'noninformative' priors yields identical results to frequentist statistics, thereby negating the advantage of Bayesian statistics. I propose instead that default choice for priors should be a weakly informative distribution that regularizes results arising from small sample sizes. I then qualitatively evaluate different priors, providing guidelines for incorporating priors into ecological analyses.

Throughout the last two decades, Bayesian statistical methods have proliferated throughout ecology and evolution (Touchon and McCoy 2016). Numerous ecology-oriented papers and textbooks established both philosophical (Ellison 2004, Clark 2005, Gelman and Shalizi 2013, Hobbs and Hooten 2015, Lemoine et al. 2016) and

© 2019 The Author. Oikos © 2019 Nordic Society Oikos

NORDIC SOCIETY OIKOS

www.oikosjournal.org

computational (Gelman and Hill 2007, Kruschke 2010, Korner-Nievergelt et al. 2015) guidelines for implementing Bayesian methods, but these sources generally emphasized the statistical model. Protocols for incorporating prior information, the defining characteristic of Bayesian philosophy, are nearly nonexistent in the ecological literature (but see Morris et al. 2015, Lemoine et al. 2016). Indeed, there has been no synthesis of the rich statistical literature on priors that describes why informative priors are preferable to noninformative priors, specifies the 'reference' priors available for common statistical models, and demonstrates how such informative priors affect posterior inference. As a result, ecologists remain understandably wary of supplementing Bayesian analyses with informative priors.

Here, I hope to encourage the use of informative priors in ecology and evolution by providing a 'consumer's guide' to weakly informative priors. This guide is structured into two sections. The first section describes why ecologists should consider using informative priors. Although the reasons outlined here are not novel, no review has synthesized the multiple disparate opinions regarding the advantages of informative priors for ecology and evolution. Specifically, I contend that ecologists should move beyond noninformative (i.e. flat, vague, diffuse) priors for three reasons: 1) in some cases, flat priors can strongly influence posterior distributions and thus are strongly informative, 2) in most cases, Bayesian analyses with noninformative priors yield identical results to computationally and theoretically simpler frequentist analyses, and 3) informative priors can mitigate type I and type M errors. The second section provides a guide for implementing informative priors, wherein I detail convenient 'reference' prior distributions for common statistical models (i.e. regression, ANOVA, hierarchical models). I then use simulations to visually demonstrate how informative priors influence posterior parameter estimates. To further encourage ecologists to use informative priors, STAN code for all models is available in appendices for each section to illustrate the simplicity of implementing informative priors.

## Moving beyond noninformative priors

Bayesian analyses differ from frequentist analyses by incorporating prior information via conditional probabilities, known as Bayes' rule:

$$\frac{\Pr(\theta)\Pr(\boldsymbol{Y}\,|\,\theta)}{\Pr(\boldsymbol{Y})} = \Pr(\theta\,|\,\boldsymbol{Y}) \tag{1}$$

where $\Pr(\theta)$ is the probability of the parameter or hypothesis $\theta$ based on prior information, $\Pr(\boldsymbol{Y}|\theta)$ is the likelihood of the data conditioned on the hypothesis, $\Pr(\boldsymbol{Y})$ is a normalization constant, and $\Pr(\theta|\boldsymbol{Y})$ is the posterior probability of the of the hypothesis conditioned on the observed data. In other words, posterior distributions describe the prior probability of the hypothesis or parameter updated with new information.

Bayesian updating proceeds, in a general sense, via weighted averaging of the prior and likelihood functions, with weights corresponding to certainties (Gelman et al. 2013). High data certainty resulting from high statistical power (i.e. large effect sizes, large sample sizes, low noise) strongly updates the prior. Likewise, highly certain priors exert a large influence on the posterior and require high-powered data to yield updating. Conversely, the posterior distribution is relatively insensitive to low certainty in either the prior (i.e. noninformative priors) or the data (i.e. low statistical power). Although the consequences of low statistical power on posterior inferences have been extensively documented (Button et al. 2013, Lemoine et al. 2016, Parker et al. 2018), relatively little attention has been paid to the consequences of noninformative priors. Indeed, ecologists often relegate prior choice to a minor concern whereby priors are chosen to stabilize computations and minimally affect the posterior (Ellison 2004, Korner-Nievergelt et al. 2015).

### Problems with noninformative priors

Ecologists typically define noninformative priors as distributions that are flat over the entire real number line and thus contain no information (Table 1). Common noninformative priors include a wide uniform distribution [e.g. $\mathcal{U}(-1000,1000)$ or $\mathcal{U}(0,1000)$ for positive-only variance parameters] or a diffuse normal distribution [e.g. $\mathcal{N}(0,10\,000)$]. Indeed, most Bayesian analyses in ecology use flat priors (Table 1). However, flatness per se does not define a noninformative prior. A $\mathcal{N}(0,4)$ distribution is noninformative if, for example, the range of plausible parameter values lies between $-1$ and 1. Conversely, a flat distribution might be strongly informative if placed on a transformed parameter. A flat $\mathcal{U}(0,1000)$ distribution is noninformative for data on small numerical scales, but highly informative distribution is the scale of the data exceeds 1000. A more accurate definition of noninformative priors would therefore be 'distributions that possess a range of uncertainty larger than any plausible parameter value' (Gelman and Hill 2007).

Assigning flat prior distributions to transformed parameters often yields highly skewed, strongly informative priors for the parameter in the original scale. Hobbs and Hooten (2015) provided the example of using flat priors to estimate a probability from binary data. The data, $\boldsymbol{y}$, are Bernoulli distributed with a probability of occurrence $p$:

$$\boldsymbol{y} \sim \text{Bernoulli}(p) \tag{2}$$

Since computational procedures struggle to estimate $p$ near the boundaries of 0 and 1, the logit transformation places $p$ on the entire real line:

$$p = \text{logit}^{-1}(a) \tag{3}$$

and the algorithm estimates $a$. In a Bayesian framework, $a$ would typically receive a flat prior [e.g. $a \sim \mathcal{N}(0,100)$] to represent complete uncertainty over the value of $p$. Yet a flat prior on $a$ heavily biases $p$ towards 0 or 1 under the diffuse normal prior on $a$ (see Fig. 5.4.3 in Hobbs and Hooten

Table 1. Number of studies from five influential ecological journals using priors of various forms. I searched every study published in all five journals from 2014 to 2018 for the term 'bayes'. Any studies professing to use 'Bayesian analyses' were searched for 'priors' to identify the type of prior used. If priors were not mentioned in the main text, these were listed as 'No mention' (including studies who may have listed priors in supplementary info). I excluded non-conventional analyses, such as phylogenetic tree construction and stable isotope mixing models. The number in parentheses is the percentage of studies falling into each category for that particular journal.

| Journal | Noninformative | Weakly informative | Strongly informative | No mention |
|---|---|---|---|---|
| Ecology | 37 (54%) | 6 (8%) | 4 (6%) | 20 (30%) |
| Ecology Letters | 11 (33%) | 2 (6%) | 0 | 20 (61%) |
| Journal of Ecology | 23 (66%) | 1 (3%) | 0 | 11 (31%) |
| Oecologia | 29 (76%) | 2 (5%) | 0 | 7 (18%) |
| Oikos | 6 (43%) | 2 (14%) | 0 | 6 (43%) |

2015). Instead, assigning $a$ the seemingly informative prior of $a \sim \mathcal{N}(0,2)$ allows for a more uniform, noninformative prior for $p$ on the original scale (see Fig. 5.4.3 in Hobbs and Hooten 2015). In a more practical example, ecologists commonly use logistic regression to estimate how $p$ changes with a given predictor $\boldsymbol{x}$:

$$\boldsymbol{p} = \text{logit}^{-1}(\boldsymbol{a}) \quad (4)$$

$$\boldsymbol{a} = \beta_0 + \beta_1 \boldsymbol{x} \quad (5)$$

Ecologists might represent complete uncertainty in $\beta_0$ and $\beta_1$ by placing diffuse normal priors on each parameter. However, such priors mistakenly place most of the distributional mass of the slope near 0 on the original scale (Fig. 1, Supplementary material Appendix 1). Symmetric distributions of the slope on the original scale are better captured by non-flat prior distributions on the logit scale (Fig. 1). Thus, ecologists should exercise extreme caution when using flat priors on transformed parameters.

Even when ecologists carefully choose prior distributions to be noninformative on the correct scale, using such noninformative priors negates the advantage of Bayesian statistics. Specifically, Bayesian estimated parameters and inferences based on noninformative priors will be identical to frequentist analyses with standard methods (e.g. the lm or glm functions in R, Gelman et al. 2013). To demonstrate, I used both frequentist and Bayesian methods to estimate parameters from a simple linear regression. I drew 10 000 estimates of each parameter $[\beta_0 \sim \mathcal{N}(-0.05, 0.15^2)$, $\beta_1 \sim \mathcal{N}(1.25, 0.5^2)$, $\log \sigma_y \sim \mathcal{N}(0.01, 0.1^2)]$ to generate 10 000 unique datasets $\boldsymbol{y}$ (each dataset contained 100 observations). Every simulated dataset was subjected to three different parameter estimation techniques: 1) ordinary least squares, 2) maximum likelihood optimization, and 3) Bayesian estimation using MCMC sampling with noninformative priors $[\beta_0, \beta_1 \sim \mathcal{N}(0, 10\,000^2)$, $\sigma_y \sim \text{InvGamma}(0.01, 0.01)]$. The standard errors for each parameter ($\sigma_\beta$) were estimated using the technique best suited to each procedure: least-squares calculations, square-root of the diagonal of the inverse Hessian matrix, and standard deviation of posterior estimates, respectively. Models were fit in Python ver. 3.6 using the statsmodels, scipy and pystan modules (Supplementary material Appendix 2).

Parameter estimates for $\hat{\beta}_1$ were equivalent to multiple decimal places for all three methods (Fig. 2A). That is, noninformative priors converged on exactly the same slope as either least-squares or maximum likelihood estimates, such that noninformative Bayesian analyses provided no unique information. Furthermore, the standard error for $\beta_1$ was nearly identical for all three methods, indicating that inferences regarding uncertainty and 'statistical significance' would also be identical for all three methods (Fig. 2B). While least-squares and Bayesian estimates of $\sigma_{\beta_1}$ fell along

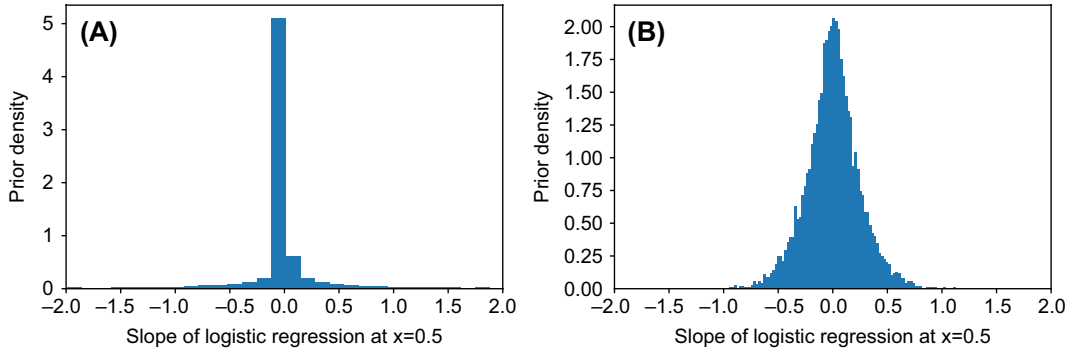


Figure 1. Noninformative priors on transformed variables can often produce strongly informative priors on the untransformed variable of interest. In (A), noninformative priors on the parameters of a logistic regression $[\beta_0, \beta_1 \sim \mathcal{N}(0, 100)]$ yield slopes that are highly spiked at 0. The panel shows the histogram of the prior distribution on the slope of a logistic regression, on the probability scale, at $x = 0.5$. In (B), weakly informative priors $[\beta_0, \beta_1 \sim \mathcal{N}(0, 2^2)]$ stabilize the prior distribution on the probability scale. The panel shows the prior distribution of the slope on the probability scale at $x = 0.5$.

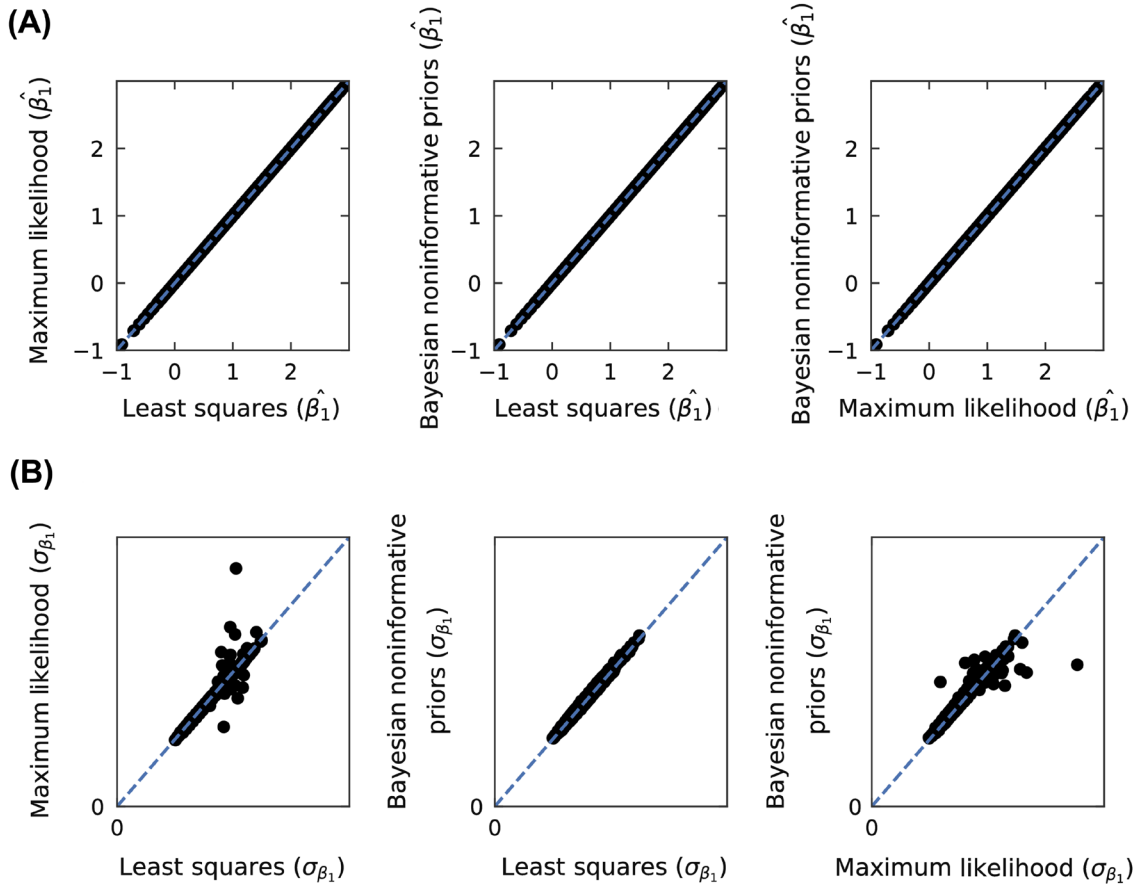


Figure 2. Least-squares, maximum likelihood, and Bayesian noninformative priors provide nearly identical estimates for (A) point estimates of the slope ($\hat{\beta}_1$) and (B) uncertainty estimates of the slope ($\sigma_{\beta_1}$) for a linear regression. Noninformative priors were $[\beta_0,\beta_1] \sim \mathcal{N}(0,10000)$ and $\sigma_y \sim \text{InvGamma}(0.01,0.01)$. The blue dashed line shows the 1:1 line of perfect equivalence. Estimates of $\sigma_{\beta_1}$ were derived from least squares methods ($\sigma_y (X'X)^{-1}$), maximum likelihood methods (inverse Hessian), or standard deviation of Bayesian posteriors. The blue dashed line shows the 1:1 line of perfect equivalence. See Supplementary material Appendix 2 for more information.

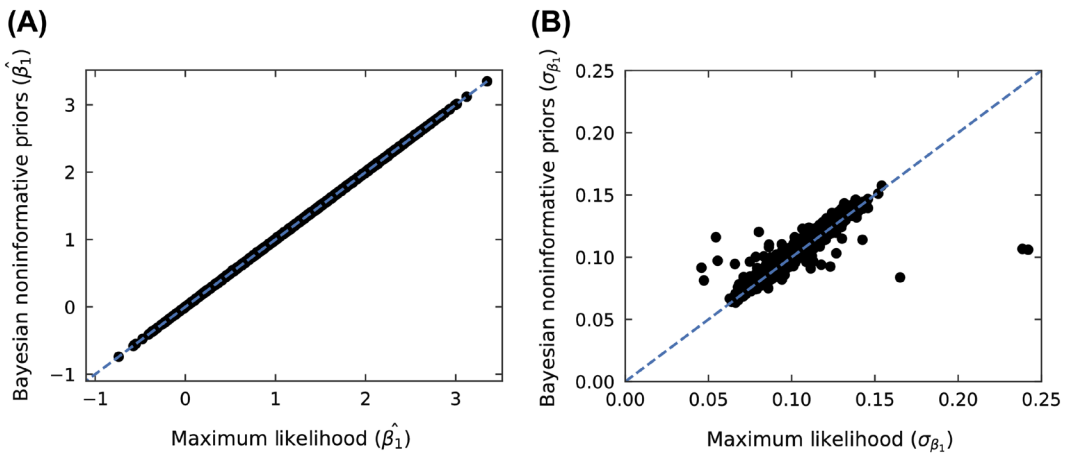


Figure 3. Maximum likelihood and Bayesian noninformative priors provide nearly identical estimates for (A) the slope ($\hat{\beta}_1$) and (B) uncertainty estimates ($\sigma_{\beta_1}$) for a mixed-effects regression. Noninformative priors were $[\beta_0,\beta_1] \sim \mathcal{N}(0,10000)$, $\sigma_a \sim \text{InvGamma}(0.01,0.01)$, $\sigma_y \sim \text{InvGamma}(0.01,0.01)$, $\mu \sim \mathcal{N}(0,10000)$, and $\alpha \sim \mathcal{N}(0,1)$. The blue dashed line shows the 1:1 line of perfect equivalence. Estimates of $\sigma_{\beta_1}$ were derived from maximum likelihood methods (inverse Hessian) or standard deviation of Bayesian posteriors. See Supplementary material Appendix 2 for more information.

the 1:1 line, maximum likelihood had a few cases where the optimizer struggled to converge to the correct estimate (Fig. 2B), suggesting that MCMC sampling is a more robust, if slower, parameter estimation method than maximum likelihood optimization for even simple linear regression models.

Just as with linear regression, parameter estimates for mixed-effects models also converge for maximum likelihood and noninformative Bayesian estimates. To illustrate, I simulated data from a multilevel model with random-effects intercepts:

$$y_j = \mu + \alpha_j + \beta_1 x_j + \epsilon_j \quad (6)$$

$$\alpha_j \sim \mathcal{N}\left(0, \sigma_\alpha^2\right) \quad (7)$$

$$\epsilon \sim \mathcal{N}\left(0, \sigma_y^2\right) \quad (8)$$

As before, I used 10 000 sets of randomly generated parameters $[\mu \sim \mathcal{N}\left(0, 0.15^2\right)$, $\beta_1 \sim \mathcal{N}\left(1.25, 0.5^2\right)$, $\alpha \sim \mathcal{N}\left(0, 0.5^2\right)$, $\log \sigma_y \sim \mathcal{N}\left(0.01, 0.1^2\right)]$ to simulate mixed-effects datasets (Supplementary material Appendix 2). For each dataset, I estimated fixed and random effects using mixed-effects models based on maximum likelihood (statsmodels module) and Bayesian methods using noninformative priors ($[\mu, \beta_1] \sim \mathcal{N}\left(0, 10\,000^2\right)$, $\sigma_a \sim$ InvGamma(0.01,0.01), $\sigma_y \sim$ InvGamma(0.01,0.01), $\alpha \sim \mathcal{N}\left(0, 1^2\right)$, pystan module). The parameters $a$ were given informative priors of $\mathcal{N}\left(0, 1^2\right)$ because of the particular modeling approach used to impart numerical stability for the MCMC sampler (Supplementary material Appendix 2). As before, estimates of $\hat{\beta}_1$ were identical between maximum likelihood and Bayesian approaches with noninformative priors (Fig. 3A). Noninformative priors also yielded identical estimates of $\sigma_{\beta_1}$, with a few exceptions where the maximum likelihood optimizer did not fully converge (Fig. 3B).

Because the same parameter estimates and statistical inferences result from both frequentist and noninformative Bayesian analyses, both methods suffer from the same statistical issues. The 'winner's curse' in particular has received considerable recent attention in both ecological and statistical literature (Button et al. 2013, Gelman and Carlin 2014, Lemoine et al. 2016). The winner's curse occurs when subsequent, repeated experiments cannot replicate an initial significant finding (Ioannidis 2005). Statistical significance of the original experiment occurred by chance, due to an overestimate of the true effect size (type M error), which caused a false positive (type I error). Type I errors are ostensibly controlled by setting $\alpha = 0.05$, but the true false positive rate (FPR) is the probability of a false positive given a significant result, which depends on statistical power: FPR = $\alpha/(\alpha + \text{power})$ (Moyé 1998, Supplementary material Appendix 3). Similarly, the true positive rate, known as the positive predictive value (PPV), is the probability of a true positive given a significant result: PPV = power/($\alpha$ + power) = 1 − FPR (Button et al. 2013, Heston and King 2017, Supplementary material Appendix 3). PPV and FPR both vary with statistical power; for studies with highly variable data, FPR can exceed 80% at low sample sizes before converging to $\alpha$ at high sample sizes (Fig. 4). In other words, with noisy data, weak effects, and small sample sizes, a statistically significant result is so unlikely that a significant result is more likely to be a false positive than a true positive. This is true regardless of whether significance is calculated by least squares, maximum likelihood, or Bayesian analyses with noninformative priors.

## Informative priors

Many authors advocate the use of informative priors to alleviate problems like the winner's curse that arise from low statistical power. Gelman and Hill (2007) and Gelman et al. (2013) state that noninformative priors facilitate model development but should be substituted for informative priors during the final analyses. More philosophically, both Kruschke (2010) and Hobbs and Hooten (2015) argue that current scientific philosophy improperly idealizes science in vacuum, wherein only the data on hand are used to draw inferences; science should instead utilize all available data. Despite these arguments, most textbooks on Bayesian data

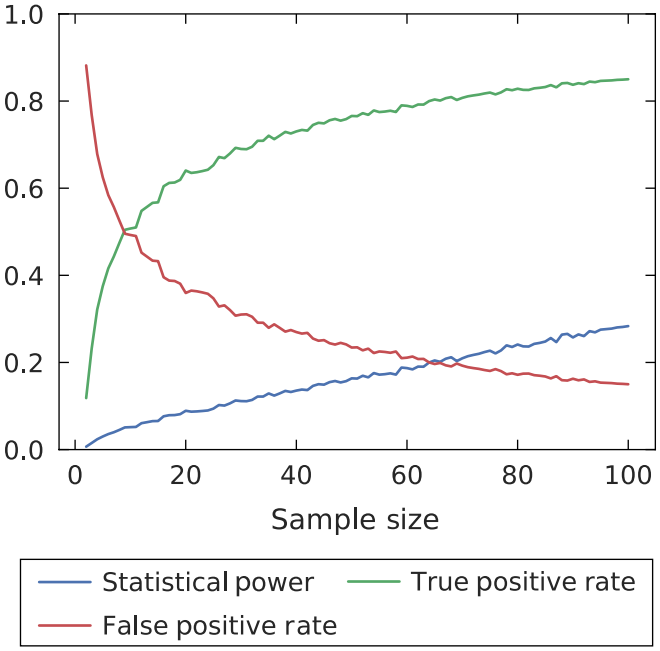


Figure 4. Small sample sizes increase the likelihood of a false positive due to low power. This figure was generated with Monte Carlo simulations of a paired t-test. The difference between groups ($\mu = 1$) was smaller than the within-group variance ($\sigma = 5$). For each simulation, I drew data for the first group [ $y_1 \sim \mathcal{N}(0, \sigma)$ ] and the second group [ $y_2 \sim \mathcal{N}(\mu, \sigma)$ ]. I then compared groups using a two-sample t-test, and significance was recorded as $p < 0.05$. This simulation was repeated 10 000 times for each sample size $N \in [2,100]$. Power at each sample size was calculated as the proportion of tests that returned significant results (true positives). I then calculated PPV and the true false positive rate using the equations defined in text. See Supplementary material Appendix 3 for more information.

analyses provide analytical and computational examples using noninformative priors, avoid making explicit recommendations for informative priors, and do not illustrate how informative priors might affect results (Gelman and Hill 2007, Kruschke 2010, Korner-Nievergelt et al. 2015, but see McElreath 2015, Lemoine et al. 2016). Because of the ambiguity surrounding prior choice and the pedagogy of Bayesian statistics emphasizing noninformative priors, informative priors remain underutilized in ecological data analysis (Table 1).

This is unfortunate because many statistical issues like the winner's curse can be mitigated via the use of informative priors. Weakly informative priors serve as a method of statistical regularization, preventing the winner's curse by shrinking parameter estimates towards zero unless there is sufficiently strong evidence of a large effect (McElreath 2015). Low-powered data result in larger shrinkage. In plainer words, large effect sizes arising from limited sample sizes and noisy data are treated skeptically. Consider a few practical examples of regularization. In the first example, imagine a t-test comparing the mean of two groups. Standardizing the response variable and using a $\mathcal{N}(0,1^2)$ prior for the difference between groups states that effect sizes are unlikely to be greater than one standard deviation of the response. This is an explicit acknowledgment that most ecological effects are small (Møller and Jennions 2002, Jennions and Møller 2003) and that unrealistically large effects in the data should be constrained unless supported by high statistical power (i.e. large sample sizes, low noise). This effectively makes Bayesian analyses with weakly informative priors more conservative, but potentially more accurate, than analyses based on noninformative priors. Weakly informative priors reduced statistical power at sample sizes below 50, but also reduced type I error rates nearly by half (Fig. 5, Supplementary material Appendix 3). As another example, I conducted 10 000 simulation experiments for an independent t-test with a statistical power of 0.2. For each simulation, I calculated the confidence interval for the effect size using both noninformative and weakly informative $\mathcal{N}(0,1^2)$ priors. Under noninformative priors, the confidence interval failed to include the true effect size in 3.59% of simulations, whereas the failure rate for weakly informative priors was only 2.87%. Shrinkage can also mitigate the need for post-hoc corrections for multiple comparisons. Gelman et al. (2012) simulated data from multiple treatment groups with small effect sizes and conducted all pairwise comparisons among groups. Out of 1000 simulations, classical tests recorded at least one significant difference 47% of the time, whereas Bayesian analyses with shrinkage recorded at least one significant difference only 5% of the time. However, Bayesian differences were 35% more accurate (i.e. correct sign) than classical analyses. Gelman et al. (2012) described the objective of regularization succinctly, as 'the price to pay for more reliable comparisons is to claim confidence in fewer of them'.

It is worth noting that the goal of regularization is agnostic of statistical philosophy and can be achieved using either frequentist (e.g. ridge, LASSO regression) or Bayesian (e.g. informative priors) methods (Gelman et al. 2017). Indeed, certain informative prior distributions in Bayesian statistics analytically reduce to frequentist ridge or LASSO estimators (Hoerl and Kennard 2000, Park and Casella 2008). However, Bayesian priors are conceptually and programmatically simpler to use, flexible in the strength of regularization, and easy to implement for generalized and hierarchical models.

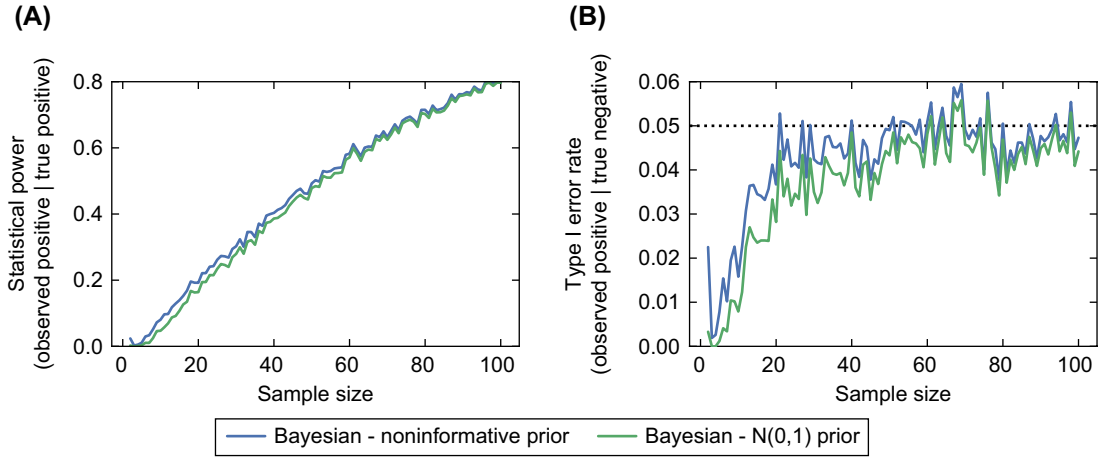


Figure 5. Weakly informative priors [ $\mathcal{N}(0,1)$ ] reduce both (A) statistical power and (B) type I error rates for t-tests compared to noninformative priors [ $\mathcal{N}(0,10\,000)$ ]. As sample size increases, prior choice has no effect on either statistical power or type I error rates. For each simulation, I drew $\boldsymbol{y}_1$ from a $\mathcal{N}(0,5)$ distribution. The effect was then either chose as false ($\mu = 0$) or true ($\mu = 1$). Then, $\boldsymbol{y}_2$ was drawn from a $\mathcal{N}(\mu,5)$ distribution. The Bayesian model estimated the difference between means ($\mu$), where $\mu$ was given either noninformative or weakly informative priors. After model fitting, I determined significance by examining whether the 95% credible interval of $\mu$ overlapped with 0. A true positive was defined as statistical significance when $\mu = 1$, and a false positive rate was defined as statistical significance when $\mu = 0$. Each rate was estimated from 10 000 simulations at each sample size. Prior to analyses, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ were combined into a single vector and standardized. Models were fit using pystan. See Supplementary material Appendix 3 for more information.

## A guide to weakly informative priors

There are several strategies for choosing informative priors. The most common perception of generating informative priors involves obtaining priors from earlier studies or expert opinion. This can be done informally, by surveying the literature and/or experts and deriving a prior distribution from those surveys (Murray et al. 2009, Martin et al. 2012) or by integrating the new data analysis within a meta-analysis of earlier studies and allowing estimates from earlier studies to serve as the prior for the new analysis (Gelman and Hill 2007, Gelman et al. 2013, Hobbs and Hooten 2015). Unfortunately, both methods are time and labor intensive. Alternatively, authors can conduct sensitivity analyses, rerunning a Bayesian analysis with progressively narrower prior distributions (Korner-Nievergelt et al. 2015). The narrowest distribution that does not affect results is thus the best 'informative' prior (Korner-Nievergelt et al. 2015). However, the sensitivity method does not identify an informative prior so much as it identifies the narrowest prior that is noninformative (see the definition of noninformative priors above). Finally, ecologists can choose a default regularization prior. Regularization priors implicitly incorporate knowledge that most ecological data are noisy and effects sizes are typically small (Møller and Jennions 2002, Jennions and Møller 2003). The statistical literature is rich with advice on regularization priors for various models (see <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations> for a thorough list of priors available for different models), and here I synthesize the prior choices available for common models and demonstrate how regularization priors affect posterior distributions.

### Linear regression models

Linear regression models have the form:

$$\hat{\boldsymbol{y}} = \boldsymbol{XB} \tag{9}$$

$$\boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{y}}, \sigma_y^2\right) \tag{10}$$

where $\boldsymbol{B}$ is a vector of regression coefficients and $\sigma_y$ is the residual standard deviation. In frequentist regression, parameters $\boldsymbol{B}$ are estimated from the data and $\sigma_y$ is a derived quantity. In Bayesian regression, both $\boldsymbol{B}$ and $\sigma_y$ are estimated parameters and thus both require priors. Typically, $\sigma_y$ can be given a noninformative prior [e.g. $\sigma_y \sim$ Cauchy(0,25)] because the emphasis of the analysis, and therefore the need for regularization, is on $\boldsymbol{B}$.

Several regularization options exist for $\boldsymbol{B}$:

1. $\boldsymbol{B} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, 1000^2\right)$

This completely noninformative prior places independent, diffuse normal priors on each parameter in the vector $\boldsymbol{B}$ (Gelman and Hill 2007, Kruschke 2010, Korner-Nievergelt et al. 2015). Parameter estimates suffer no regularization and will be identical to estimates from frequentist regression (Fig. 2, 3).

2. $\boldsymbol{B} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, 1^2\right)$

This prior states that parameter estimates will most likely fall between $-1$ and 1, and almost always fall between $-2$ and 2 (McElreath 2015, Lemoine et al. 2016). Because this prior is a standard normal distribution, ecologists must carefully consider the scale of both the predictor and response. For example, this prior can be noninformative when the units of $\boldsymbol{y}$ are very small (e.g. 0.001–0.009 g) or strongly informative when the units of $\boldsymbol{X}$ are very small (i.e. the change in $\boldsymbol{y}$ per unit change in $\boldsymbol{X}$ will be large). The safest option for using this prior is to standardize both the response and predictors, such that the prior states that a one standard deviation change in the predictor is unlikely to yield more than a 1 standard deviation change in the response and will almost never yield more than a 2 standard deviation change in the response. This prior reduces to frequentist ridge regression (Hoerl and Kennard 2000).

3. $\boldsymbol{B} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, 0.5^2\right)$

This prior requires that $\boldsymbol{y}$ is also standardized to $\mathcal{N}(0, 0.5)$, and is conceptually similar to the above prior (Gelman et al. 2008). This prior is more useful for logistic regression.

4. $\boldsymbol{B} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, 1/\tau_B\right),\ \tau_B \overset{i.i.d.}{\sim} \text{Gamma}\left(\alpha, \beta\right)$

The normal-gamma prior places an independent normal distribution on each parameter, and the inverse variance of each normal distribution is modeled as an independent gamma distribution. The normal-gamma prior is the completely continuous version of the discrete spike-and-slab prior and, under certain $\alpha$ and $\beta$ priors, similar in shape to the Laplace distribution (Griffin and Brown 2010, 2017). Like the Laplace distribution, the normal-gamma distribution spikes sharply near zero, thus providing stronger regularization than the $\mathcal{N}(0,1)$ prior. This prior also requires careful consideration of the scale of both the response and predictors, and is a generalized version of frequentist LASSO regression (Park and Casella 2008, Griffin and Brown 2010). Careful choice of $\alpha$ and $\beta$ allow the prior to become more or less informative (Supplementary material Appendix 4).

5. $\boldsymbol{B} \sim \mathcal{N}\left(0, \sigma_B^2\right),\ \sigma_B \sim \text{Cauchy}(0, \nu) \in [0, \infty]$

This prior models $\boldsymbol{B}$ hierarchically, treating the parameters in $\boldsymbol{B}$ as coming from a positive-truncated Cauchy distribution with some unknown variance among parameters. By modeling $\boldsymbol{B}$ hierarchically, results from each predictor informs the others (Kruschke 2010). Since most predictors will have weak parameter estimates, this prior shrinks all parameters towards zero. The degree of shrinkage depends on $\nu$, which controls the width of the Cauchy prior on $\sigma_B$. An noninformative prior on $\sigma_B$ with large $\nu$ yields a standard mixed-effects model with random parameters (Gelman 2006). While standard mixed effects models do provide regularization, using

a small value of $\nu$ implies that there should be little variation among parameters and imparts stronger regularization (Supplementary material Appendix 4).

These priors work for models that contain only quantitative predictors or a mixture of quantitative and categorical predictors. Categorical predictors must be encoded with caution. Dummy coding (0,1) treats a change in category as a unit change in $\boldsymbol{X}$, such that the $\mathcal{N}\left(0,1^2\right)$ prior, for example, states that changing the category or treatment level should not result in more than a 2 standard deviation change in the predictor. Effects coding, on the other hand, represents each factor as a deviation from the overall data mean. A $\mathcal{N}\left(0,1^2\right)$ prior then assumes that treatments should not cause a large deviation from the mean. The special case of models that contain only categorical predictors (i.e. ANOVA designs) is discussed below.

Simulation experiments of a linear regression with an intercept (B0) and four predictors (B1, B2, B3, B4) demonstrated the regularizing effect of weakly informative priors (Supplementary material Appendix 4). Noninformative priors often produced type M errors when coupled with noisy data (Fig. 6). Parameters B1, B2 and B3 were, for example, all overestimated at $N = 10$ and $\sigma_y = 8$. At $N = 20$, 30 and $\sigma_y = 8$, parameter B4 exhibited similar type M errors with noninformative priors (Fig. 6). Weakly informative priors mitigated type M errors by constraining parameter estimates closer to the true value for small effects at the cost of occasionally underestimating large true effects (Fig. 6). For example, at $N = 10$ and $\sigma_y = 8$, regularizing priors prevented overestimation of B1 and B2 while shrinking estimates of B3 closer to the true value. Similarly, at $N = 20$, 30 and $\sigma_y = 8$, regularizing priors shrunk erroneously large estimates of B4 towards zero and the true value. However, shrinkage also caused weakly informative models to underestimate the true effect of B3 at $N = 20$ and $\sigma_y = 8$. At $\sigma_y = 2$, the regularizing effect of weakly informative priors weakened until, at large sample sizes, prior choice had no impact on model results. Thus, ecologists should consider adopting a normal-gamma or hierarchical weakly informative prior as a default prior for regressions; such priors conservatively estimate parameters and mitigate type M errors with low-powered data but have little impact on posterior inference with high-powered data.

## ANOVA models

ANOVA models are a special case of linear regression comprised of only categorical predictors. In a hierarchical ANOVA, categorical factors are grouped into $k$ batches of coefficients corresponding to different factors or sets of interaction terms (Gelman 2005, Gelman and Hill 2007). As an example, suppose an ecologist initiates an experiment to assess whether aquatic primary production is independently or jointly regulated by nitrogen (N) and phosphorus (P) (Allgeier et al. 2011). Each nutrient treatment has two levels (0, 10 g m$^{-2}$), resulting in four treatments (N0–P0, N10–P0, N0–P10, and N10–P10). This experiment can be analyzed with a hierarchical two-way ANOVA:

$$\boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{y}}, \sigma_y^2\right) \tag{11}$$

$$\hat{\boldsymbol{y}} = \mu + \boldsymbol{B}_N + \boldsymbol{B}_P + \boldsymbol{B}_{NP} \tag{12}$$

$$\boldsymbol{B}_N \sim \text{Prior} \tag{13}$$

$$\boldsymbol{B}_P \sim \text{Prior} \tag{14}$$

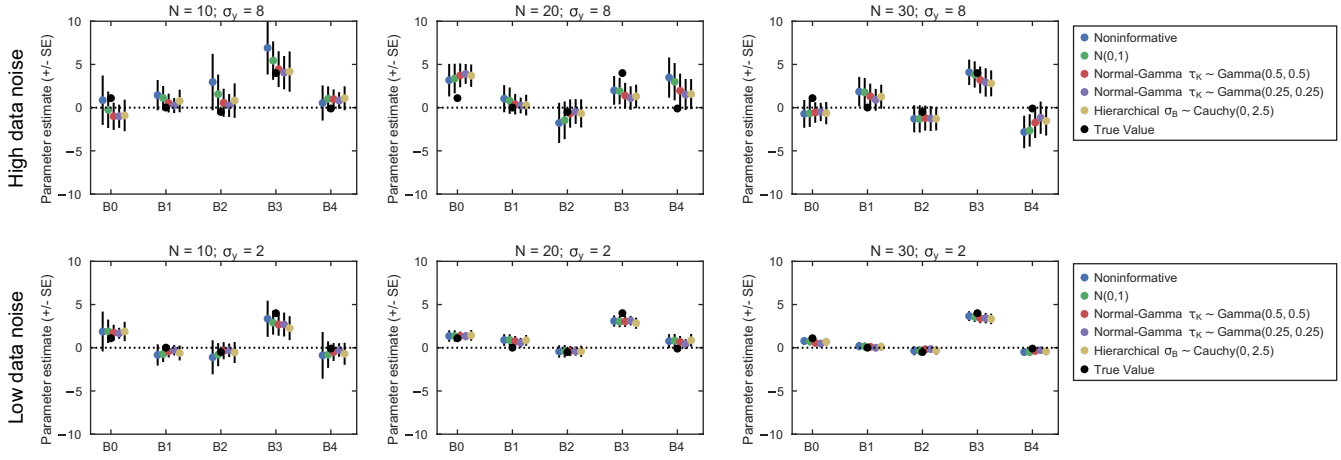


Figure 6. A single example of how weakly informative regularizing priors affect posterior inference for parameters B0, B1, B2, B3 and B4 in a multiple regression. Regressions were run at various sample sizes ($N = 10$, 20, 30) and data noise ($\sigma_y = 2$, 8). Note that the dataset was fixed within each panel, such that the same dataset was analyzed with all five models. However, the dataset varied among panels. Points represent the best posterior point estimate, bars are one standard error. Supplementary material Appendix 4 contains all code for simulation exercises. Thorough demonstrations of how priors affect shrinkage based on Monte Carlo simulations or analytical results are found in Lemoine et al. (2016) – Supplementary material Appendix 2 Fig. A3 and Gelman et al. (2012) – Fig. 4.

$$\boldsymbol{B}_{NP} \sim \text{Prior} \tag{15}$$

Here, μ is the overall mean, $\boldsymbol{B}_N$ contains the marginal means of each N treatment, $\boldsymbol{B}_P$ contains the marginal means of each P treatment, and $\boldsymbol{B}_{NP}$ contains the cell means of the joint N–P treatments. Since ANOVAs emphasize comparing between- and within-groups variance components, this model is purposefully overparameterized to estimate variance components for the marginal and cell means.

Because this model is over-parameterized and non-identifiable, linear constraints are imposed on each batch of coefficients to induce identifiability:

$$\boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{y}}, \sigma_y^2\right) \tag{16}$$

$$\hat{\boldsymbol{y}} = \mu^* + \boldsymbol{B}_N^* + \boldsymbol{B}_P^* + \boldsymbol{B}_{NP}^* \tag{17}$$

$$\mu^* = \mu + \mathrm{E}\left[\boldsymbol{B}_N\right] + \mathrm{E}\left[\boldsymbol{B}_P\right] + \mathrm{E}\left[\boldsymbol{B}_{NP}\right] \tag{18}$$

$$\boldsymbol{B}_N^* = \boldsymbol{B}_N - \mathrm{E}\left[\boldsymbol{B}_N\right] \tag{19}$$

$$\boldsymbol{B}_P^* = \boldsymbol{B}_P - \mathrm{E}\left[\boldsymbol{B}_P\right] \tag{20}$$

$$\boldsymbol{B}_{NP}^* = \boldsymbol{B}_{NP} - \mathrm{E}\left[\boldsymbol{B}_{NP}\right] \tag{21}$$

$$\boldsymbol{B}_N \sim \text{Prior} \tag{22}$$

$$\boldsymbol{B}_P \sim \text{Prior} \tag{23}$$

$$\boldsymbol{B}_{NP} \sim \text{Prior} \tag{24}$$

The parameters $\boldsymbol{B}^*$ now represent marginal and cell mean deviations from the overall mean (see Supplementary material Appendix 5 for other potential ANOVA model formulations). Finite population variation among groups can be calculated from the posteriors of each batch of coefficients (Gelman 2005):

$$\mathrm{Var}[N] = \mathrm{Var}\left[\boldsymbol{B}_N - \mathrm{E}\left[\boldsymbol{B}_N\right]\right] = \mathrm{Var}\left[\boldsymbol{B}_N^*\right] \tag{25}$$

$$\mathrm{Var}[P] = \mathrm{Var}\left[\boldsymbol{B}_P - \mathrm{E}\left[\boldsymbol{B}_P\right]\right] = \mathrm{Var}\left[\boldsymbol{B}_P^*\right] \tag{26}$$

$$\mathrm{Var}[NP] = \mathrm{Var}\left[\boldsymbol{B}_{NP} - \mathrm{E}\left[\boldsymbol{B}_{NP}\right]\right] = \mathrm{Var}\left[\boldsymbol{B}_{NP}^*\right] \tag{27}$$

$$\epsilon = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{28}$$

$$\mathrm{Var}[\text{Residual}] = \mathrm{Var}[\epsilon] \tag{29}$$

Comparing the relative sizes of each variance component provides an estimate not only of the significance but also, and perhaps more importantly, the importance of each factor. However, ecologists should be aware of three important differences between hierarchical and traditional ANOVAs. First, classical ANOVAs are sensitive to unbalanced designs and must be conducted with the appropriate sums-of-squares, while implementation of hierarchical ANOVAs does not depend on balanced data. Second, mean-square-error calculations for each factor in a classical ANOVA assume that every other variance component except the residual is zero. In hierarchical ANOVAS, variance components are estimated simultaneously and thus will differ from classical estimates (Gelman 2005). Finally, factors are considered 'statistically significant' in classical ANOVAs only when the variance component (mean-square-error) is acceptably larger than residual error. In a hierarchical ANOVA, significant factors need not have larger variance components than residual error; indeed, they often do not. However, comparing factor variance components to residual error provides an estimate of importance for each factor conceptually similar to $\omega^2$ from classical ANOVAs (Bruno et al. 2006).

As variance components are estimated from the variation among treatment effects, regularization priors that shrink treatment effects ($\boldsymbol{B}$) towards the overall mean will also regularize the variance components. Several regularization options exist and many are similar to the priors for regression (see Supplementary material Appendix 5 for all model code).

1. $\boldsymbol{B}^{(k)} \overset{i.i.d.}{\sim} \mathcal{N}\left(\mu^{(k)}, 1000^2\right)$

This prior gives every coefficient in the *k*th batch an independent, diffuse prior and provides no shrinkage. However, estimates of variance components will differ from frequentist analyses because variance components are estimated simultaneously rather than independently. In my personal experience, this model regularly fails to converge.

2. $\boldsymbol{B} \overset{i.i.d.}{\sim} \mathcal{N}\left(\mu^{(k)}, 1^2\right)$

This prior states that deviations for any treatment or interaction term should be <2, though the scale of the response variable needs to be carefully considered. Note that because $\boldsymbol{B}$ contains the means for each treatment in the batch, the parameters $\boldsymbol{B}$ need not be centered around 0, but are instead centered around the mean of the two groups $\mu^{(k)}$. Thus, this prior states that treatment means should not deviate from the mean of treatment means by more than 2. If $\boldsymbol{y}$ has been standardized, then $\mu^{(k)}$ can be replaced with 0, because the overall mean will be 0.

3. $\boldsymbol{B}^{(k)} \sim \mathcal{N}\left(\mu^{(k)}, \sigma_B^{2(k)}\right),\ \sigma_B^{2(k)} \sim \mathrm{Cauchy}\left(0, \nu^{(k)}\right) \in [0, \infty]$

This prior models each batch of coefficients hierarchically, treating each batch of coefficients *k* as coming from a

normal distribution with a batch-specific standard deviation. As $\nu$ increases, the prior on $\sigma_B^{2(k)}$ becomes essentially uniform and parameter estimates converge to the same parameter estimates obtained from maximum likelihood mixed-effect models (Gelman 2005). Because variances are difficult to estimate when the number of groups is small (i.e. variance among two nitrogen treatments), informative priors of $\sigma_B^{2(k)} \sim \text{Cauchy}(0, 2.5)$ help constrain variance components to realistic values (Gelman 2006).

4. $\mathbf{B}^{(k)} \sim \mathcal{N}\left(\mu^{(k)}, \sigma_B^{2(k)}\right),\ \sigma_B^2 \sim \text{Cauchy}(0, \nu) \in [0, \infty]$

This prior is slightly different from the above prior. Instead of modeling each variance component $\sigma_B^{2(k)}$ as coming from an independent prior distribution, all variance components are modeled with a single, hierarchical distribution (Gelman 2005). This prior states that most batches should have little variation among coefficients, but allows for occasional large variance components. Stronger priors restrict the probability of large variance components and thus provide more shrinkage.

The severity of regularization induced by weakly informative priors depends not just on statistical power but also on data balance. With balanced data, weakly informative priors constrained estimates for treatments that deviated substantially from the overall mean and did so most strongly for low-powered data (N = 5, $\sigma_y = 8$, Fig. 7). Treatment estimates for high-powered data, resulting from either low noise ($\sigma_y = 2$) or larger sample sizes (N = 30) exhibited less regularization with weakly informative priors (Fig. 7). In contrast, frequentist ANOVAs always yielded treatment estimates that converged to the data mean regardless of statistical power (Fig. 7). Unbalanced data led to more regularization for all models. Indeed, constraining estimates of the unbalanced treatment to the overall mean is one of the lesser appreciated aspects of frequentist ANOVAs. In contrast, hierarchical ANOVAs with weakly informative priors regularized all treatment means, especially those that deviated significantly from the overall mean (Fig. 7). Some degree of regularization of treatment means occurred across all permutations of sample size and data noise except in the ideal case of high sample sizes (N = 30) coupled with low noise ($\sigma_y = 2$). Thus, weakly informative priors constrained treatment means only for less ideal situations: unbalanced data and/or low power.

## Hierachical models

Bayesian statistics have become synonymous with hierarchical models. These models treat some combination of parameters as arising from an 'overall' or 'population-level' distribution that allows extrapolation to unobserved cases. For example, Belenky et al. (2003) measured the reaction reaction time of humans to subjected to sleep deprivation for ten consecutive days (sleepstudy data available in the R lme4 package). A standard linear model for these data has the form:

$$\hat{y}_j = \beta_0 + \beta_1 \text{Day} + \beta_{2j} \text{Subject}_j + \beta_{3j} \text{Day} \times \text{Subject}_j \quad (30)$$

$$y \sim \mathcal{N}\left(\hat{y}, \sigma_y^2\right) \quad (31)$$

where $\beta_0$ is the intercept of subject 1, $\beta_1$ is the slope of reaction time per day for subject 1, $\beta_{2j}$ is the difference in baseline reaction time between subject 1 and subject *j*, and $\beta_{3j}$ is the difference in slopes between subject 1 and subject *j*. Here, subjects are treated as 'fixed-effects' and, although the intercept and slopes are estimated for each observed subject, extrapolation of intercepts and slopes to the general population of unobserved individuals is impossible.

Extrapolation requires the use of hierarchical, or random-effects, models. A random-intercept model accounts for the non-independence of measurements by allowing intercepts to vary among subjects:

$$\hat{y}_j = \alpha_j + \beta_1 \text{Day}_j \quad (32)$$

$$\alpha_j \sim \mathcal{N}\left(\mu_\alpha, \sigma_\alpha^2\right) \quad (33)$$

$$y \sim \mathcal{N}\left(\hat{y}, \sigma_y^2\right) \quad (34)$$

where $\alpha_j$ is the intercept of the *j*th subject, $\mu_\alpha$ is the population-level intercept, and $\sigma_\alpha^2$ is the between-subject variance in $\alpha$. This model is a repeated-measures design (in R, repeated measures designs are fit in the lme4 package by treating subject as random: reaction ~ time + (1|subject). This R formula corresponds to the random-intercept model here.) that assumes all subjects share the same slope. This assumption can be relaxed by treating slopes as a second random variable (in R, this model is fit in lme4 with the formula reaction ~ time + (time|subject). The lme4 model will assume that $\alpha$ and $\gamma$ are correlated.):

$$\hat{y}_j = \alpha_j + \gamma_j \text{Day}_j \quad (35)$$

$$\alpha_j \sim \mathcal{N}\left(\mu_\alpha, \sigma_\alpha^2\right) \quad (36)$$

$$\gamma_j \sim \mathcal{N}\left(\mu_\gamma, \sigma_\gamma^2\right) \quad (37)$$

$$y_j \sim \mathcal{N}\left(\hat{y}, \sigma_y^2\right) \quad (38)$$

Intercepts and slopes now have distributions that enable researchers to extrapolate results to the general population. The shape of these distributions is controlled by the variance parameters, which in a Bayesian context receive priors. Weakly informative priors regularize population

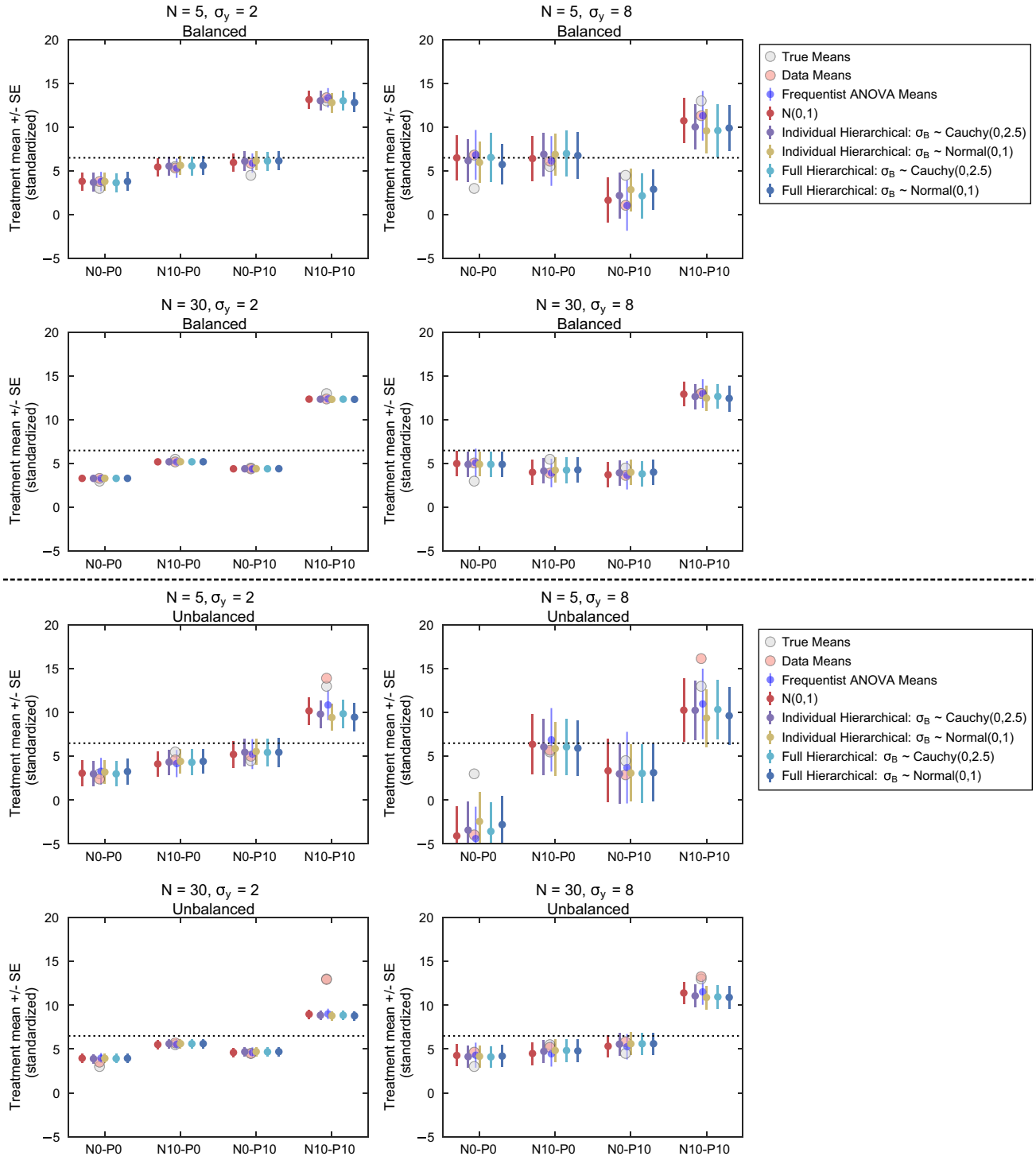


Figure 7. A single example of how weakly informative regularizing priors affect posterior inference in a two-way ANOVA. N refers to the number of replicates per treatment, i.e. N = 5 means that each of the four treatments contained five observations. To create unbalanced data, I reduced the number of replicates in the N10–P10 treatment by half. Points represent the best posterior point estimate, bars are one standard error. Supplementary material Appendix 5 contains all code for simulation exercises.

variances and, in so doing, shrink the individual parameter estimates.

***Priors for random-intercept models***

Priors for random-intercept models should regularize $\sigma^2_\alpha$, reducing the among-subject spread of intercepts unless supported by a large number of subjects. Importantly, the number of subjects now dictates parameter shrinkage, not the number of data points within each subject (see Supplementary material Appendix 6 for each prior distribution).

1. $\log\sigma^2_\alpha \sim \mathcal{U}(-100,100)$

This prior is noninformative for $\sigma^2_\alpha$ on the log-scale, but highly informative on the natural scale (Browne and Draper 2006, Gelman and Hill 2007). On the original scale, this prior gives exponentially more weight to larger variances and, since log0 is undefined, results in an infinite probability mass at $\sigma^2_\alpha = 0$. Thus, this prior is not recommended (Browne and Draper 2006).

2. $\sigma_\alpha \sim \mathcal{U}(0,100)$

This prior is completely noninformative and reduces to maximum-likelihood estimation of between-subject variances (Gelman and Hill 2007). Note that this prior is placed on the standard deviation, rather than the variance.

3. $\sigma^2_\alpha \sim \mathcal{U}(0,1/e)$

This prior is also completely noninformative but, unlike the previous distribution, places the prior on the variance not the standard deviation. It is highly sensitive to the choice of $e$. Typically, $e = 0.001$ (Browne and Draper 2006).

4. $\sigma^2_\alpha \sim \text{Gamma}(1.5,10^{-4})$

This prior is weakly informative and skewed towards zero (Chung et al. 2015). Note that $10^{-4}$ is the rate (scale$^{-1}$) of the gamma distribution.

5. $\sigma^2_\alpha \sim \text{InvGamma}(e,e)$

This prior should be used carefully. It becomes improper as $e$ decreases and can inflate variances from small sample sizes (Daniels 1999).

6. $\sigma_\alpha \sim \text{Cauchy}(0,\nu)\epsilon[0,\infty]$

This prior is as above for regression and ANOVA models. McElreath (2015) suggests setting $\nu = 1$.

I analyzed the sleepstudy data with random-intercept hierarchical models using each of the six weakly informative priors listed above (see Supplementary material Appendix 6 for model code). Because priors were placed on $\sigma^2_\alpha$, as opposed to the subject intercepts, I demonstrate shrinkage using histograms of posterior distributions for $\sigma_\alpha$ (see Supplementary material Appendix 6 for plots of subject intercepts under each prior).

The histograms and confidence intervals in Fig. 8 showcase several important aspects of weakly informative hierarchical priors. First, $\log\sigma^2_\alpha \sim \mathcal{U}(-100,100)$ is unacceptable because the prior mass on $\sigma^2_\alpha$ is so infinitesimally small as to eliminate low among-subject variance from the posterior regardless of sample size (Fig. 8). Across all sample sizes, the lower limit of $CI_{95}$ was bounded at 1. Second, uniform priors did not shrink posterior distributions and yielded unrealistically high estimates ($CI_{95} > 29$) at $N = 3$ (Fig. 8). Conversely, the InvGamma(0.1,0.1) prior regularized priors too harshly by placing too much posterior mass on $\sigma_\alpha \approx 0$. Third, the two Cauchy priors provided the best posteriors for $\sigma_\alpha$ at $N = 3$, with Cauchy(0,1) reducing the right tail slightly more than Cauchy(0,2.5) ($CI_{95} = 3.84$, 4.91 respectively, Fig. 8). Finally, all priors converged to a stable posterior distribution as sample size increased (Fig. 8). At $N = 9$, the InvGamma(0.1,0.1) and two Cauchy priors provided more shrinkage than other priors, evidenced by the shorter right tail and smaller $CI_{95}$ for $\sigma_\alpha$ (Fig. 8). At $N = 18$, all proper posterior distributions were essentially identical, with a slight degree of increased shrinkage for the InvGamma(0.1,0.1) and two Cauchy priors compared to other priors (Fig. 8).

***Priors for random-intercept and random-slope models***

The random-intercept model assumed that the relationship between reaction time and duration of sleep deprivation is constant among all subjects. Exploratory data analysis, on the other hand, indicated substantial variability in slopes among subjects (Supplementary material Appendix 6). It might therefore be more appropriate to account for among-subject variation in slopes. The model in Eq. 35–38, as described in Kruschke (2010), assumes that intercepts and slopes are independent. In practice, random effects should be modeled as correlated variables (Gelman and Hill 2007):

$$\hat{\boldsymbol{y}}_j = \boldsymbol{X}_j\boldsymbol{B}_j \tag{39}$$

$$\boldsymbol{B} \sim \mathcal{MVN}\left(\mu_B,\Sigma_B\right) \tag{40}$$

$$\boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{y}},\sigma^2_y\right) \tag{41}$$

where $\boldsymbol{X}$ is the design matrix containing an intercept and duration data, $\boldsymbol{B}_j$ is a vector containing subject-specific intercepts and slopes, $\mu_B$ is a vector of the population-level intercept and slope, and $\Sigma_B$ is the covariance matrix for intercepts and slopes. Regularization priors must constrain variances and so are placed on $\Sigma_B$.

1. $\Sigma_B \sim \text{InvWish}(\boldsymbol{I},k+1)$

The inverse Wishart prior requires two parameters: a $k \times k$ identity matrix $\boldsymbol{I}$ and $k+1$ degrees of freedom, where $k$ is the number of random effects. Using $k+1$ degrees of freedom is equivalent to setting a uniform distribution from $-1$ to 1 on

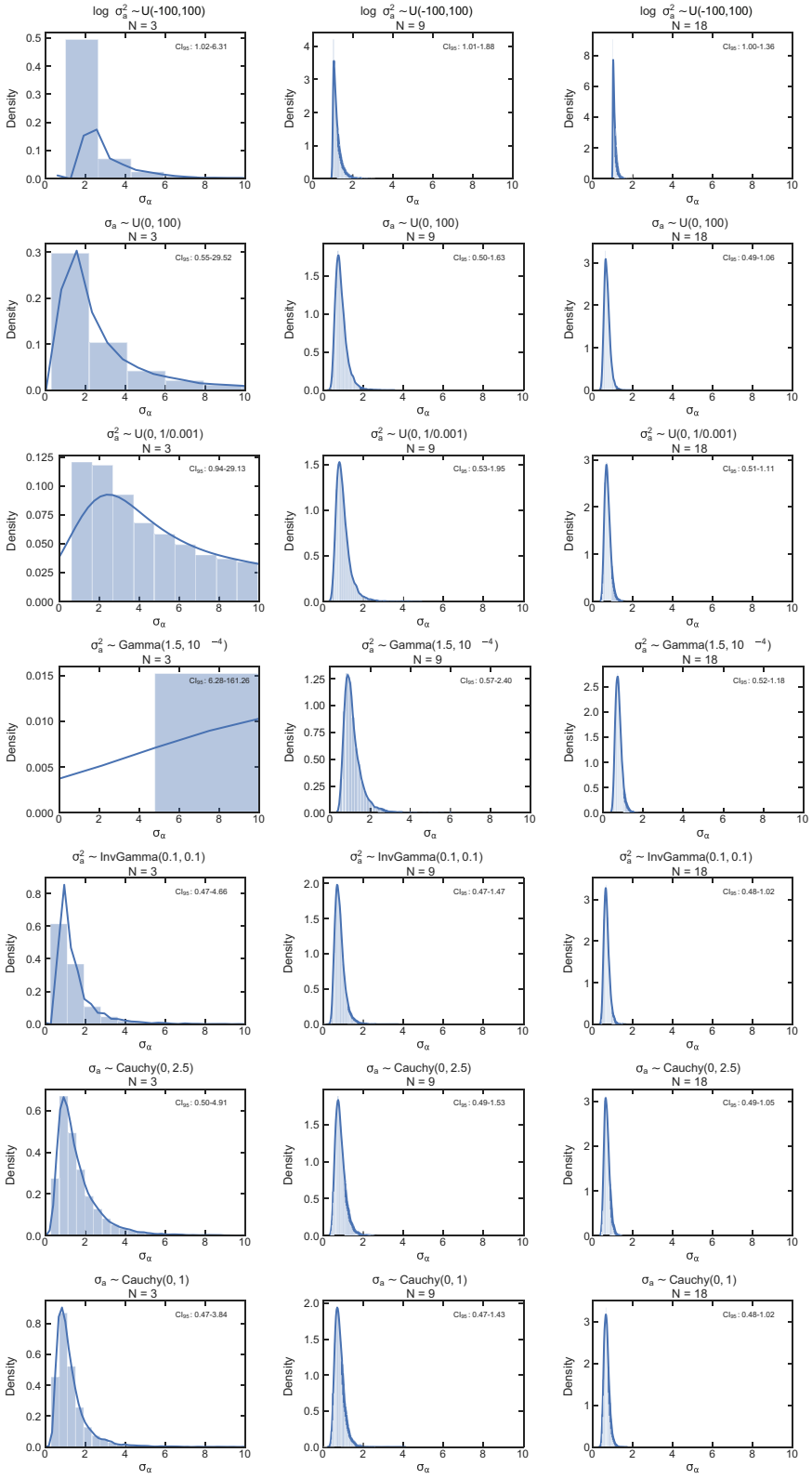


Figure 8. Histograms of posterior distributions for $\sigma_\alpha$ from a single example of a random-intercept model of the sleepstudy data. Sample sizes were varied by taking the first N subjects. Supplementary material Appendix 6 contains all code.

correlations among parameters (Gelman and Hill 2007, Alvarez et al. 2014). The InvWish prior gained popularity as the conjugate prior for a multivariate normal distributions, facilitating computations via Gibbs sampling in BUGS and JAGS. However, the InvWish prior performs poorly when variances are small relative to means, constraining variances upwards and correlations downwards (Alvarez et al. 2014).

2. $\Sigma_B = \text{Diag}(\tau)\Omega\text{Diag}(\tau)$, $\tau \sim \mathcal{U}(0,100)$, $\Omega \sim \text{InvWish}(\boldsymbol{I}, k+1)$

An alternative to the InvWish prior is scaled inverse Wishart prior (Gelman and Hill 2007). This prior still uses the inverse Wishart distribution to maintain conjugacy but alleviates the issue of bias by decomposing $\Sigma_B$ into two components: a vector of scaling coefficients for variances $\tau$ and an unscaled covariance matrix $\Omega$. Note that $\Omega$ is not the correlation matrix and $\tau$ and $\Omega$ cannot be interpreted independently. The scaled InvWish prior alleviates bias by estimating variances without constraint via $\tau$ while placing a uniform prior on correlations in $\Omega$. However, both $\tau$ and $\Omega$ receive noninformative priors, providing no regularization.

3. $\Sigma_B = \text{Diag}(\tau)\Omega\text{Diag}(\tau)$, $\tau \sim \text{Cauchy}(0,\nu) \in [0,\infty]$, $\Omega \sim \text{LKJ}(\phi)$

Since STAN does not require conjugacy, $\Sigma_B$ can be decomposed into conceptually simpler components. Now, $\tau$ is a vector containing the variances of each parameter and can be given a Cauchy prior for shrinkage as discussed for random-intercept models, and $\Omega$ is the correlation matrix among parameters. The LKJ distribution has one parameter $\phi$. When $\phi = 1$, the prior is uniform over all possible correlation matrices. As $\phi > 1$, the prior becomes more heavily spiked towards the identity matrix thereby shrinking correlations among parameters, and as $\phi < 1$ the prior becomes biased against the identity matrix (Stan Development Team 2016). Weakly informative priors use $\phi$ slightly greater than one.

4. $\Sigma_B \sim \text{Wish}\left(\left(1/2\times10^{-4}\right)\boldsymbol{I}, k+2\right)$

The covariance matrix is given a Wishart prior with the identity matrix $\boldsymbol{I}$ multiplied by a large number and degrees of freedom equal to the number of parameters $k+1$. This prior arises from the product of Gamma(1.5,$10^{-4}$) priors on the eigenvalues of $\Sigma_B$ (Chung et al. 2015).

Re-analysis of the sleepstudy data using the model of Eq. 39–41 revealed marked differences in posterior distributions of $\sigma_\alpha$ and $\sigma_\gamma$ among the four priors from $\Sigma_B$ (see Supplementary material Appendix 7 for model code). The InvWish prior resulted in prior posteriors but biased estimates upwards for $\sigma_\gamma$, as indicated by the slightly higher bounds for the $CI_{95}$ compared to other priors (Fig. 9). I therefore follow Alvarez et al. (2014) in recommending that ecologists avoid the InvWish prior. Under the scaled InvWish and Wish priors, posteriors for $\sigma_\alpha$ and $\sigma_\gamma$ were improper at N = 3 (Fig. 9). At N = 3, only the two Cauchy priors resulted in reasonable posteriors for both the intercept and slope. As sample size increased, the scaled InvWish and Cauchy priors converged on stable posterior distributions, albeit with slightly more shrinkage present in the Cauchy(0,1) posterior at N = = 9 (Fig. 9). I recommend that ecologists use the Cauchy priors when modeling correlated random effects for four reasons: 1) the decomposed components are easily interpretable, $\tau$ contains variances and $\Omega$ is the correlation matrix, 2) the Cauchy alone yielded proper posteriors across all sample sizes, 3) priors can be easily be strengthened or weakened via Cauchy and LKJ distributions, or by replacing Cauchy(0,1) with $\mathcal{N}(0,1)$, and 4) Cauchy posteriors converged towards the noninformative scaled InvWish prior at high sample sizes.

## Concluding thoughts

In this manuscript, I have argued that ecologists should, at a minimum, incorporate weakly informative priors in Bayesian analyses in order to stabilize posteriors and regularize parameter estimates. Regularization can potentially mitigate or eliminate common errors associated with frequentist or noninformative Bayesian analyses, and weakly informative priors are simple to implement in any standard Bayesian model (Supplementary material Appendix 1–7). In this concluding section, I offer two final thoughts on practicing Bayesian statistics with weakly informative priors.

First, weakly informative priors provide modest, not severe, regularization. Readers may have noticed throughout the numerous examples provided here that posterior distributions from weakly informative priors almost always overlap with noninformative posteriors (Fig. 6–9). Indeed, the difference between noninformative priors and weakly informative priors would generally not be 'statistically significant'. The distributional overlap among priors does not discredit the need for priors but instead is the hallmark of a well-chosen weakly informative prior. Strongly informative priors would, on the other hand, significantly affect posterior distributions and are the subject of other reviews (Murray et al. 2009, Martin et al. 2012, Morris et al. 2015). Even the modest regularization provided by weakly informative priors with low statistical power can reduce type I errors (Fig. 5) and improve out-of-sample prediction for regression models (McElreath 2015). In ANOVA models, shrinkage provided by regularization priors can eliminate the need for post-hoc corrections for multiple comparisons; pairwise contrasts of weakly informed posteriors often have lower error rates than Bonferroni-corrected pairwise comparisons among noninformed posteriors (Gelman et al. 2012). With high statistical power, differences among priors cease. The safest course of action is therefore to begin with a weakly informative prior to regularize low-powered data and have no effect on high-powered data.

Second, given the variety of available priors and the ease with which multiple priors can be implemented, ecologists should always conduct Bayesian analyses utilizing multiple prior distributions of varying strength. Once an initial model has been coded and embedded within a pre- and

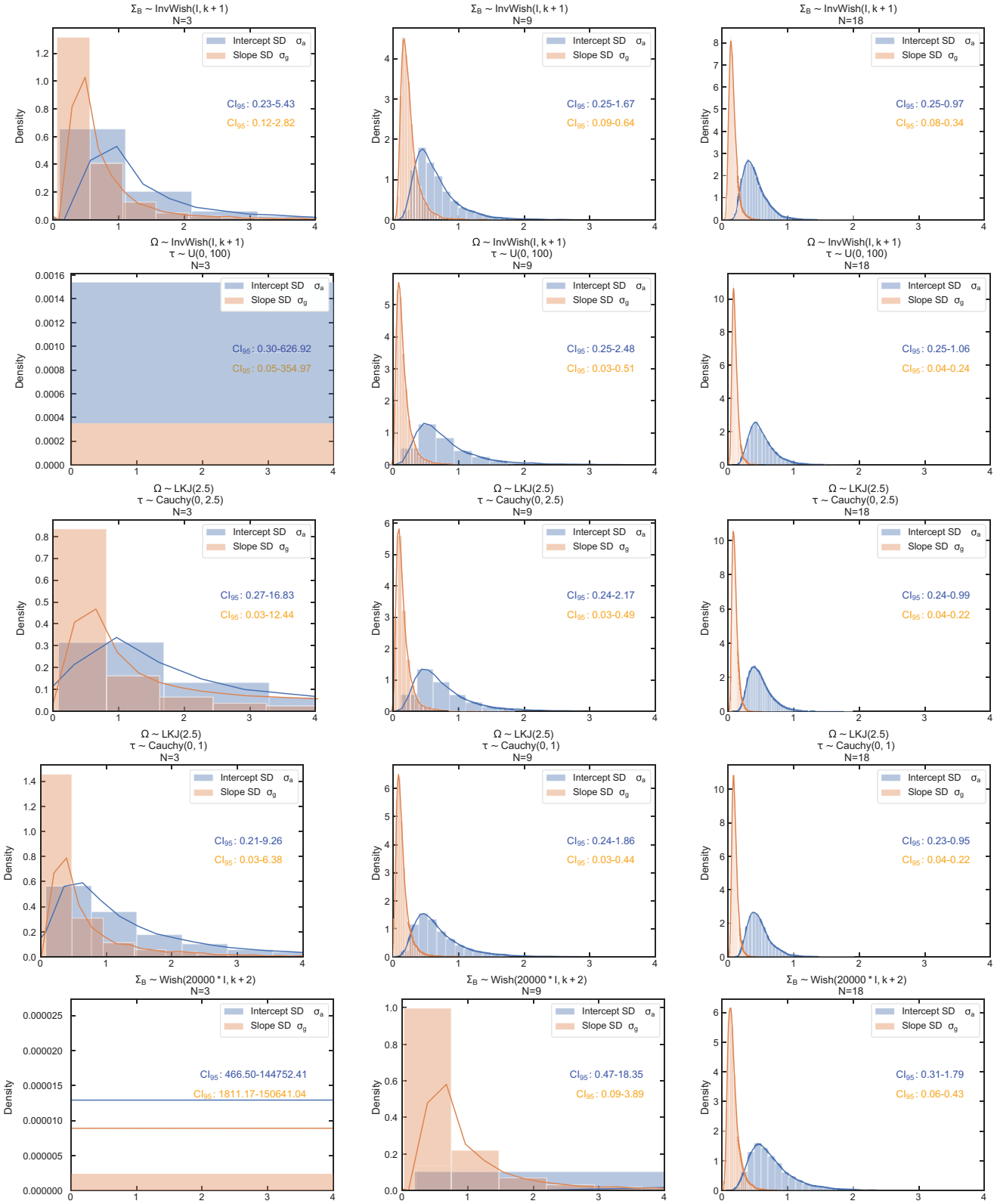


Figure 9. Histograms of posterior distributions for $\sigma_\alpha$ and $\sigma_\gamma$ from a single example of a random-intercept and random-slope model of the sleepstudy data. Sample sizes were varied by taking the first N subjects. Supplementary material Appendix 7 contains all code.

post-processing script, it is trivial to copy the script and alter priors to generate new figures or embed multiple models within the same processing script. Regardless of approach, data should be analyzed using multiple priors along a gradient of information. A completely noninformative prior allows ecologists to judge patterns in the data alone. Comparing noninformative posteriors to progressively stronger priors informs the ecologist about the strength of patterns. If posteriors are sensitive to prior choice then the ecologist should be cautious in their interpretation and choose a regularized prior for presentation as the main analysis. Conversely, posteriors that do not differ among priors enable ecologists to make strong inferences from their data. Posteriors from multiple priors can be presented together in the main text (i.e. Fig. 7, Rode et al. 2017) or as supplementary materials. By comparing results among multiple priors, ecologists will be explicitly forced to consider the power of their data and strength of effect sizes. During my survey of the literature, a surprising number of papers did not report prior information in the main text or anywhere in the supplemental literature. Papers using Bayesian methods should always report the priors used and the justification for each prior (see Rode et al. 2017 for an example). This information is crucial to interpretation of the results and should be present in the main text.

Finally, I would be remiss if I did not provide concrete recommendations for prior choice. For linear regressions, I recommend the normal-gamma prior due to its similar shrinkage with the hierarchical prior and flexibility in altering the strength of priors (Fig. 6). For example, priors for interaction terms can be strengthened as interactions should require more evidence, or interaction terms can be modeled hierarchically such that a weak prior for the interaction requires that one or both main effects be strong (Griffin and Brown 2017). Likewise, the full hierarchical priors for ANOVA models provide the best balance of regularization, ease of implementation, and flexibility (Fig. 7). For both random-intercept and random slope models, ecologists should use Cauchy priors because only Cauchy priors provided proper posteriors at low sample sizes. Other priors biased estimates upwards or were improper at low sample sizes. Of course, prior choice is subjective, but ecologists should never use priors with hard limits (i.e. $\mathcal{U}(-1000,1000)$), even for variance parameters. Variances should instead be modeled with an unbounded, positive-only distribution.

In summary, Bayesian analyses depend on prior information. Priors should not be considered a nuisance component of Bayesian analyses whose only role is to stabilize posteriors, but rather constitute an integral part of Bayesian statistical philosophy, interpretation, and model fitting. With the guidelines provided here, I hope to normalize the use of weakly informative priors for Bayesian analyses in ecology. Ecologists can and should debate the appropriate form of prior information, but should consider weakly informative priors as the new 'default' prior for any Bayesian model.

### Alternative viewpoints

During the review process, I received several comments from reviewers (both friendly and anonymous) that Bayesian methods are not only used to incorporate priors, but also to fit complex models (e.g. hierarchical hurdle models, models with custom distributions, missing data models) that have no least squares solution and fail under maximum likelihood optimization. Such models only work using Bayesian MCMC parameter estimation. However, I believe that MCMC samplers are independent of Bayesian analyses for two reasons. First, it is possible to fit models via MCMC sampling without priors. Figure 2 and 3 demonstrate that MCMC samplers can yield the same estimates as OLS, and are more robust than maximum likelihood. Second, it is possible to conduct Bayesian analyses without using MCMC samplers. Bayesian linear regression and some simple hierarchical models have analytical solutions under conjugate priors, such that MCMC samplers are not necessary (Gelman et al. 2013). Yet because analytical solutions do not exist for many Bayesian models, MCMC samplers have become synonymous with Bayesian statistics. It is my opinion that ecologists carefully distinguish between the use of MCMC samplers for model fitting and the practice of Bayesian statistics, which requires priors.

*Acknowledgments* – I would like to thank A.A. Shantz, D.E. Burkepile, J.D. Parker and J. Byrnes for comments on earlier drafts. *Funding* – National Science Foundation, Directorate for Biological Sciences, Division of Environmental Biology, 1754124. US Dept of Agriculture, Natl Inst. of Food and Agriculture 2016-67012+25169.

## References

Allgeier, J. E. et al. 2011. The frequency and magnitude of non-additive responses to multiple nutrient enrichment. – J. Appl. Ecol. 48: 96–101.

Alvarez, I. et al. 2014. Bayesian inference for a covariance matrix. – arXiv:1408.4050.

Belenky, G. et al. 2003. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose–response study. – J. Sleep Res. 12: 1–12.

Browne, W. J. and Draper, D. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. – Bayesian Anal. 1: 473–514.

Bruno, J. F. et al. 2006. Partitioning the effects of algal species identity and richness on benthic marine primary production. – Oikos 115: 170–178.

Button, K. S. et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. – Nat. Rev. Neurosci. 14: 365–376.

Chung, Y. et al. 2015. Weakly informative prior for point estimation of covariance matrices in hierarchical models. – J. Educ. Behav. Stat. 40: 136–157.

Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. – Ecol. Lett. 8: 2–14.

Daniels, M. J. 1999. A prior for the variance in hierarchical models. – Can. J. Stat. 27: 567–578.

Ellison, A. M. 2004. Bayesian inference in ecology. – Ecol. Lett. 7: 509–520.

Gelman, A. 2005. Analysis of variance – why it is more important than ever. – Ann. Stat. 33: 1–53.
Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). – Bayesian Anal. 1: 515–534.
Gelman, A. and Carlin, J. B. 2014. Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. – Perspect. Psychol. Sci. 9: 641–651.
Gelman, A. and Hill, J. 2007. Data analysis using regression and multilevel/hierarchical models. – Cambridge Univ. Press.
Gelman, A. and Shalizi, C. R. 2013. Philosophy and the practice of Bayesian statistics. – Br. J. Math. Stat. Psychol. 66: 8–38.
Gelman, A. et al. 2008. A weakly informative default prior distribution for logistic and other regression models. – Ann. Appl. Stat. 2: 1360–1383.
Gelman, A. et al. 2012. Why we (usually) don't have to worry about multiple comparisons. – J. Res. Educ. Effect. 5: 189–211.
Gelman, A. et al. 2013. Bayesian data analysis, 3rd edn. – Chapman and Hall.
Gelman, A. et al. 2017. The prior can often only be understood in the context of the likelihood. – Entropy 19: 1–13.
Griffin, J. E. and Brown, P. J. 2010. Inference with normal-gamma prior distributions in regression problems. – Bayesian Anal. 5: 171–188.
Griffin, J. E. and Brown, P. J. 2017. Hierarchical shrinkage priors for regression models. – Bayesian Anal. 12: 135–159.
Heston, T. F. and King, J. M. 2017. Predictive power of statistical significance. – World J. Methodol. 7: 112–116.
Hobbs, N. T. and Hooten, M. B. 2015. Bayesian models: a statistical primer for ecologists. – Princeton Univ. Press.
Hoerl, A. E. and Kennard, R. W. 2000. Ridge regression: biased problems for nonorthogonal estimation. – Technometrics 42: 80–86.
Ioannidis, J. P. A. 2005. Why most published research findings are false. – PLoS Med. 2: e124.
Jennions, M. D. and Møller, A. P. 2003. A survey of the statistical power of research in behavior ecology and animal behavior. – Behav. Ecol. 14: 438–445.

Supplementary material (available online as Appendix oik-05985 at <www.oikosjournal.org/appendix/oik-05985>). Appendix 1–7.

Korner-Nievergelt, F. et al. 2015. Bayesian data analysis in ecology using linear models with R, BUGS and STAN. – Academic Press.
Kruschke, J. K. 2010. Doing Bayesian data analysis: a tutorial with R and BUGS, 1st edn. – Academic Press.
Lemoine, N. P. et al. 2016. Underappreciated problems of low replication in ecological field studies. – Ecology 97: 2554–2561.
Martin, T. G. et al. 2012. Eliciting expert knowledge in conservation science. – Conserv. Biol. 26: 29–38.
McElreath, R. 2015. Statistical rethinking: a Bayesian course with examples in R and Stan. – Chapman & Hall/CRC Press.
Møller, A. P. and Jennions, M. D. 2002. How much variance can be explained by ecologists and evolutionary biologists? – Oecologia 132: 492–500.
Morris, W. K. et al. 2015. The neglected tool in the Bayesian ecologist's shed: a case study testing informative priors' effect on model accuracy. – Ecol. Evol. 5: 102–108.
Moyé, L. A. 1998. P-value interpretation and alpha allocation in clinical trials. – Ann. Epidemiol. 8: 351–357.
Murray, J. V. et al. 2009. How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies Petrogale penicillata. – J. Appl. Ecol. 46: 842–851.
Park, T. and Casella, G. 2008. The Bayesian LASSO. – J. Am. Stat. Assoc. 103: 681–686.
Parker, T. H. et al. 2018. Empowering peer reviewers with a checklist to improve transparency. – Nat. Ecol. Evol. 2: 929–935.
Rode, M. et al. 2017. Prospective evidence for independent nitrogen and phosphorus limitation of grasshopper (*Chorthippus curtipennis*) growth in a tallgrass prairie. – PLoS One 12: e0177754.
Stan Development Team 2016. Stan: a C++ library for probability and sampling, ver. 2.14 pages – <http://mc-stan.org/>.
Touchon, J. C. and McCoy, M. W. 2016. The mismatch between current statistical practice and doctoral training in ecology. – Ecosphere 7: e01394.