Specifying Prior Distributions

Prior distributions for unknown parameters in a Bayesian statistical model are an important element of Bayesian modeling. They are essential and something that must be chosen, just as the likelihood is. In principle, priors serve as our existing baseline knowledge before we observe the data that we seek to analyze. It is important to be thoughtful when manifesting existing scientific knowledge in the form of a prior, but doing so is more consistent with the scientific process and can improve inference.

In principle, we should not specify priors in a way that depends on the data we seek to analyze. That is, the prior needs to be based on scientific information that exists before we see the new data to guarantee the resulting posterior distribution will be valid. However, so-called empirical Bayes methods use the data to help specify some of the parameters or hyperparameters in advance. While not truly Bayesian, empirical Bayes methods seem to work well in practice and some non-Bayesians use them for inference in some cases.

It is important to recognize efforts focused on "objective" Bayesian inference. While no prior can be specified to be truly non-informative, the phrase is still often used to describe flat or vague priors. Many objective Bayesian developments have focused on providing default priors to use in various settings that have certain properties like invariance to transformation (see Jeffreys priors below). One good use for vague priors is when testing code and algorithms before they are fit to real data. Of course, we can also use informative priors along with simulated data to assess how much leverage the prior has on the resulting inference (i.e., a sensitivity analysis). An easy way to robustify the inference relative to the prior is to collect more data. In situations where less data are available, we expect (and would want) the prior to affect the inference.

By contrast, the concept of using a prior as a constraint to reduce model complexity and thereby improve predictive ability is prevalent in non-Bayesian statistics where it is often referred to as "regularization" (e.g., Hooten and Hobbs, 2015). Penalized likelihood methods such as ridge regression and lasso are good examples of regularization. Bayesian priors can be used in the same capacity, to shrink or penalize certain parameters in ways that lead to better model performance for a given purpose.

First-moment parameters (like location parameters, e.g., μ) are relatively robust to prior choices, but higher-order parameters, often having more complicated support, are more difficult to estimate and thus are more sensitive to the prior.

Conjugacy (matching the form of the prior and posterior or full-conditional; more below) can be good for computation by allowing us to retain every MCMC realization and it does not require tuning. However, there is no rule that says conjugate priors are more appropriate in general for a given model. Choose the prior that best expresses the existing scientific information.

For non-conjugate priors, be sure to use a proposal distribution for the parameter in the same transformation as in the prior. If the transformation of the parameter is different in the proposal than the prior, the Metropolis-Hastings ratio needs to be corrected using a change of variables technique (or alternative).

Specifying informative hyperparameters in a prior for a specific data analysis can be done in a variety of ways (Hobbs and Hooten, 2015) discuss a variety of methods for example. For some parameters, it might be sufficient to set the location of the prior based on the expert knowledge. When more knowledge exists, we can specify the mean and variance of the prior and then back-calculate the natural hyperparameters in the prior using "moment-matching." An alternative is to specify the priors based

on its upper and lower quantiles so that it covers the range of realistic values we would expect before the data are collected.

Simulation from the data model itself based on a variety of candidate values for parameters can often provide insights that help us specify informative priors. For example, if certain parameter values lead to simulated data sets that we know are completely unrealistic, then we could narrow or shift the prior to avoid those parameter values.

A Few Specific Priors

When specifying a prior, the first thing to consider is the support of the variable being modeled. For example, for a variance component, we know it can only be positive, thus we could use one of the positive support distributions (e.g., gamma or inverse gamma) or specify the prior on a transformation of the variable that expands the support (e.g., $\log(\sigma) \sim \mathrm{N}(\mu_\sigma, \tau_\sigma^2)$).

Consider a few of the priors often described in the literature. Each of these have advantages and disadvantages.

- **Conjugate Prior:** A prior distribution whose form is same as that of the posterior or full-conditional. They are more difficult to derive, but if tractable, can yield automatic Gibbs sampling algorithms which require no tuning. Conjugacy does not guarantee that chains will mix well, and they may be highly informative for some parameters (e.g., the inverse gamma for Gaussian variance components). For many parameters they can be chosen to be "vague," "diffuse," or only weakly informative. Conjugate priors are most commonly used for mean parameters and regression coefficients and, for these purposes, they work well.

- **Uniform Prior:** $[\theta] \propto 1$. The uniform prior may be attractive to some because it is "flat," but the caveat is that it is not flat for transformations of the parameter. Consider the binomial-beta model. If one uses a $\mathrm{Beta}(1,1)$ distribution for the prior on the probability parameter then it is a uniform prior, and $[\theta] \propto 1$. However, suppose we're interested in θ^2, then (using the Jacobian transformation) we will end up with an implicit prior on $[\theta^2]$ that is not proportional to one (actually, $[\theta^2] \propto \frac{1}{2}(\theta^2)^{-\frac{1}{2}}$). Thus, the values of θ^2 under the prior are not equally probable as they are for θ. A potentially bigger problem with uniform priors are that they are improper (i.e., integrate to 1) if the support isn't enforced. Some improper priors can lead to a proper posterior, but that property has to be checked.

- **Jeffreys Prior:** $[\theta] \propto I(\theta)^{\frac{1}{2}}$, where $I(\theta)$ is the Fisher information (i.e., $I(\theta) = -E_y\left(\frac{d^2 \log[\mathbf{y}|\theta]}{d\theta^2}\right)$). One can show that this prior is invariant to transformation. Unfortunately, Jeffreys priors are often called "non-informative," but they are really just "reference" priors; something to use when we don't know what to use but like their properties. It can be shown that the Jeffreys prior for the binomial-beta model is $\mathrm{Beta}(0.5, 0.5)$. Notice that Jeffreys priors depend on the likelihood (kind of like conjugate priors). They can sometimes be found analytically or numerically, but not always.

- **Variance Component Priors:** If the data model is normal, then an inverse gamma is conjugate for the variance parameter σ^2. As an alternative to the inverse gamma distribution for a prior on Gaussian variance components, a log-normal seems to perform well (i.e., $\log(\sigma) \sim \mathrm{N}(\mu_\sigma, \tau_\sigma^2)$). It is not conjugate, but we could use Metropolis-Hastings updates for this parameter in an MCMC algorithm, and it allows you to specify the prior mean and variance for $\log(\sigma)$ directly. Aim for a mean close to what you might expect from similar data and then avoid making the prior variance huge. Alternatively, Gelman (2006) recommends trying a uniform distribution for σ in hierarchical models for the "process" variance component (among other things). This seems to work well only when the dimensionality of the process is at least moderately sized (i.e., ≥ 8).