

Introduction to Bayesian Statistics with WinBUGS

Part 4—Priors and Hierarchical Models

Matthew S. Johnson

New York ASA Chapter Workshop
CUNY Graduate Center
New York, NY
hspace1in
December 17, 2009

December 17, 2009

On selecting a prior

- ▶ To this point we have been discussing Bayesian concepts without really discussing where the prior distribution comes from.
- ▶ The introduction of the prior distribution into the model is probably the most controversial aspect of Bayesian statistics.
- ▶ Frequentists see the introduction of the prior distribution as biasing the end results with data-less information. Indeed, in any problem different prior distributions can lead to different conclusions.
- ▶ The best way to think about the prior distribution is that it is part of the overall model that is being used. It should be assessed whenever possible.

Different philosophies for selecting priors

- ▶ The prior distribution reflects the uncertainty about the parameter value prior to observing the data. Some authors have suggested using data defined prior distribution, but that's really double-dipping into the information provided by the data.
- ▶ The classical Bayesian views the prior distribution as a necessary tool, with which it is possible to make probability statements about the unknown parameters.
The choice of prior distribution should be “automatic” and the final results should depend only minimally on the prior distribution used.
The priors utilized by the classical Bayesian are often called flat, non-informative, or reference priors.

Examples of “flat” priors

The improper uniform prior

Consider a uniform distribution $U(-\tau, \tau)$ with density

$$f_{\theta}(\theta) = \frac{1}{2\tau} I_{\{\theta \in (-\tau, \tau)\}}.$$

The prior does not “prefer” any value θ_0 over any other value θ_1 within the range $(-\tau, \tau)$, because the probability that $\theta \in (a, b)$ depends only on $b - a$, i.e.,

$$P(\theta \in (a, b)) \propto b - a. \quad (1)$$

Now suppose we force $\tau \rightarrow \infty$, while keeping the property specified in (1), then we end up with a prior *measure* $f_{\theta}(\theta) \propto 1$.

This *measure* is not a probability measure because it not integrable. However, it can lead to a proper posterior distribution (i.e., the posterior integrates to one). Some authors state, “the prior is proportional to Lebesgue measure” when using this prior distribution.

Examples of “flat” priors

Jeffreys' prior

One of the shortcomings of the uniform prior distribution is that it is not invariant to transformations of the parameter.

For example, if one researcher assumes a uniform prior on the parameter θ , and another researcher assumes a uniform prior on the transformed parameter $\psi = \exp\{\theta\}$, then the two analyses are not necessarily equivalent.

Jeffreys' prior,

$$f_{\theta}(\theta) \propto \sqrt{J(\theta)},$$

where $J(\theta)$ is the Fisher information for θ , ensures that posterior distribution based on Jeffreys' priors are invariant to one-to-one transformations of the parameter θ .

Jeffreys' prior is often improper (i.e., it does not integrate to one), and therefore, may lead to an improper posterior.

More on “flat” priors

Many authors suggest that the arguments behind the use of the uniform and Jeffreys' priors are often overstated.

- ▶ When do we really have *no* information about a parameter value?
- ▶ If we have a reasonable amount of data then the results should be relatively robust to the choice of (reasonable) prior distributions.

Why should we worry about invariance to transformations, and propriety of the posterior distribution if we don't need to?

Parametric Bayes and Conjugate Priors

- ▶ The parametric Bayesian utilizes common probability distributions to model the uncertainty about the parameter value.
- ▶ In the same way that we often model the sampling distribution with simple distributions like the normal, exponential, or Poisson distributions, the parametric Bayesian assumes convenient distributions for f_{θ} .
- ▶ Probably the simplest class of prior distributions is the class of *conjugate* prior distributions.
- ▶ A prior distribution is conjugate for the sampling distribution if the prior and posterior distributions are members of the class of prior distributions, e.g., if both the prior and posterior distributions are normal distributions.

Another Example

Suppose that Y_1, \dots, Y_n are the responses of n individuals to a question about the performance of President Bush. We assume that $Y_i \sim \text{Ber}(\theta)$. Hence the likelihood is:

$$f_{\mathbf{y}|\theta}(\mathbf{y} | \theta) = \theta^{y_+} (1 - \theta)^{n - y_+}$$

If we assume a $\text{Beta}(a, b)$ prior distribution for the proportion θ

$$f_{\theta}(\theta) \propto \theta^{a-1} (1 - \theta)^{b-1}$$

then we have the posterior density

$$f_{\theta|\mathbf{y}}(\theta | \mathbf{y}) \propto \theta^{a+x_+-1} (1 - \theta)^{b+n-x_+-1},$$

which is the density of a $\text{Beta}(a + x_+, b + n - x_+)$ random variable. Hence the family of beta distributions is conjugate for the Bernoulli distribution.

- ▶ Conjugate priors simplify computation and are relatively easy to interpret.
- ▶ Non-conjugate priors are as justifiable as conjugate priors, the only difference is that they may complicate computations.
- ▶ In multi-parameter problems it is often difficult to find a multidimensional conjugate prior so conditional conjugate priors are utilized (i.e., those where the prior and the conditional posterior are from the same class of distributions). For example,
 - ▶ the conjugate prior for μ conditional on σ in the $N(\mu, \sigma^2)$ distribution is a normal distribution, i.e., for a normal prior on μ , the conditional posterior for μ given σ is also normal.
 - ▶ the conjugate prior for σ^2 conditional on μ in a $N(\mu, \sigma^2)$ distribution is an inverse-gamma distribution.
- ▶ Conjugate priors are especially useful in MCMC, because most common distributions have well-established algorithms to draw pseudo-random variables.
- ▶ Conjugate and all parametric priors do impose some structure on the model. That assumed structure should be evaluated.

Conjugate priors for various distributions

Sampling Distribution	Parameter ₁ Parameter ₂	Conjugate Prior
$N(\mu, \sigma^2)$	$\mu \sigma$	Normal
$N(\mu, \sigma^2)$	$\sigma^2 \mu$	Inverse-Gamma
Bernoulli(θ)	θ	Beta
Binomial(n, θ)	θ	Beta
Poisson(λ)	λ	Gamma
Exponential(λ)	λ	Gamma
Multinomial(θ)	θ	Dirichlet

Consider a model with a multidimensional parameter

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$$

- ▶ We often assume (prior) independence of the parameters (and observations) out of convenience.
- ▶ These assumptions are often unreasonable:
 - ▶ The same subject responds to several questions.
 - ▶ Students are clustered within schools.
 - ▶ The same judge examines several papers.

We must have a way to include the dependence structure.

- ▶ One way to incorporate the dependence is with a *Bayesian hierarchical* model.

The hierarchical Bayes model decomposes the prior density $f_{\theta}(\boldsymbol{\theta})$ into several conditional densities

$$f_1(\boldsymbol{\theta} | \eta_1), f_2(\eta_1 | \eta_2), \dots, f_m(\eta_{m-1} | \eta_m)$$

and a marginal density $f_{m+1}(\eta_m)$.

The resulting *marginal* prior distribution for the multidimensional parameter $\boldsymbol{\theta}$ is:

$$f_{\theta}(\boldsymbol{\theta}) = \int_{\eta_1 \times \dots \times \eta_m} f_1(\boldsymbol{\theta} | \eta_1) \cdots f_m(\eta_{m-1} | \eta_m) f_{m+1}(\eta_m) d\eta_1 \cdots d\eta_m$$

- ▶ We call η_j the hyperparameter(s) at level j .
- ▶ Normally we would like to assume $\boldsymbol{\theta}$ is conditionally independent given η_1 , so that

$$f_1(\boldsymbol{\theta} | \eta_1) = \prod_i f(\theta_i | \eta_1)$$

- ▶ η_j could be multidimensional itself, and similar conditional independence assumptions could be made.
- ▶ Hierarchical models can be considered as a special class of graphical models.

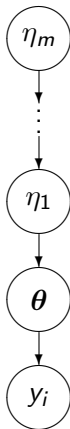
The Directed Acyclic Graph

The developers of WinBUGS suggest that the analyst starts with the graphical representation of the structural model assumptions. The structural model specifies all observable variables and all unobservable parameters and how these quantities are related. All model quantities are represented by *node* in the graph.

WinBUGS separates the nodes into three classes.

- ▶ *Stochastic nodes.* All random quantities in the model are represented by stochastic nodes. In our setup y_1, \dots, y_n , $\theta_1, \dots, \theta_k$, and $\eta_1, \dots, \eta_{m+1}$.
- ▶ *Logical nodes.* All quantities that are simply functions of other parameters, e.g., $\mu_j = \alpha + \beta x_j$, is a function of the quantities α , β and x_j .
- ▶ *Constant nodes.* These represent all quantities that are fixed by the experiment, e.g., regressors, treatments, etc.

A generic DAG



Developing the DAG

1. Start with a single node representing a single observation from distribution/random mechanism you're trying to model.
2. Ask yourself how many parameters are needed to describe the random mechanism behind that observation, e.g.
 - ▶ If y_i is continuous, then maybe you need a mean and variance.
 - ▶ If y_i is multinomial, then you need a parameter vector with the right number of probabilities.

Create the node(s) for these parameter(s) and directed edges going from the parameters to the observable.

3. Decide whether the parameter nodes should be logical (deterministic) or stochastic.
4. Develop the hierarchy above the level 1, level 2, ..., level m parameters in the same fashion.

DoodleBUGS Demonstration

Example

I gave the students in my introductory statistics class three quizzes and two exams this past semester. When developing the model to analyze the data I would like to consider the following:

- ▶ Scores for a single student should be related across tasks.
- ▶ The mean and standard deviations of scores will vary across tasks.
- ▶ The average difficulties of the exams (quizzes) will be related.

Go to WinBUGS