Steven L. Scott
University of Southern California
Bridge Hall 401H
Los Angeles, CA 90089-1421
Phone: (213) 740-8009
Fax: (213) 740-7313
email: sls@usc.edu
word count: $\sim$ 754

Prior Distribution. A prior distribution $p(\theta)$ is a probability distribution describing one's subjective belief about an unknown quantity θ before observing evidence in a data set $\mathbf{x}$. Bayesian inference combines the prior distribution with the likelihood of the data to form the posterior distribution used make inferences for θ. There are four philosophies concerning the prior distribution's role in Bayesian inference. The proponents of these philosophies are the subjectivists, the objective Bayesians, the regularizers, and the modelers.

At is foundations, Bayesian inference is a theory for updating one's subjective beliefs about θ upon observing $\mathbf{x}$. Subjectivists use probability distributions to formalize an individual's beliefs. Each individual's prior distribution is to be elicited by considering his willingness to engage in a series of hypothetical bets about the true value of θ. Models are required in order to make prior elicitation practical for continuous parameter spaces. Because of their computational convenience, *conjugate* priors are often used when they are available. A model has a conjugate prior if the prior and posterior distributions belong to the same family. Table 1 lists conjugate likelihood-prior relationships for several members of the exponential family. The prior parameters in many conjugate prior-likelihood families may be thought of as prior data. This provides an obvious way to measure the strength of the prior (e.g. "one observation's worth of prior information"). See then entry on POSTERIOR DISTRIBUTION for an example calculation using conjugate priors.

[Table 1 about here]

Very often one finds that any reasonably weak prior has a negligible effect on the posterior distribution. Yet counter examples exist, and critics of the Bayesian approach find great difficulty in prior elicitation (Efron, 1986). The *objective Bayesians* answer this criticism by deriving "reference priors" which attempt to model prior ignorance in standard situations. Kass and Wasserman (1996) review the vast literature on reference priors. Many reference priors are improper (i.e. they integrate to ∞). For example, a common "non-informative" prior for mean parameters is the uniform prior $p(\theta) \propto 1$, which is obviously improper if the parameter space of θ is unbounded. Improper priors pose no difficulty so long as the likelihood is sufficiently well behaved for the posterior distribution to be proper, which usually happens in simple problems where frequentist procedures perform adequately. However the propriety of the posterior distribution in complicated models can be difficult to check (Hobert and Casella, 1996).

One difficulty with reference priors is that noninformative priors on one scale become informative after a change of variables. For example, a uniform prior on $\log \sigma^2$ becomes $p(\sigma^2) \propto 1/\sigma^2$ because of the Jacobian introduced by the log transformation. *Jeffreys' priors* are an important family of reference prior that are invariant to changes of variables. The general Jeffreys' prior for a model $p(\mathbf{x}|\theta)$ is $p(\theta) \propto \det(J(\theta))^{1/2}$ where $J(\theta)$ is the Fisher information from a single observation.

Prior distributions are often used to regularize the parameters of complicated models (Hastie *et al.*, 2001). For example in the regression model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I_n)$, where $\mathbf{X}$ is the $n \times p$ design matrix, the prior $\beta \sim \mathcal{N}(0, \tau^2 I_p)$ leads to a posterior distribution for β equivalent to ridge regression,

$$p(\beta|\mathbf{X}, \mathbf{y}, \tau, \sigma^2) = \mathcal{N}(B, \Omega^{-1}).$$

The posterior precision (inverse variance) $\Omega = (\mathbf{X}^T\mathbf{X}/\sigma^2 + I_p/\tau^2)$ is the sum of the prior precision $\Omega_0 = I_p/\tau^2$ and the likelihood precision $\Omega_1 = \mathbf{X}^T\mathbf{X}/\sigma^2$. If $\hat{\beta}$ is the least squares estimate of β then the posterior mean B can be written

$$B = (\Omega_0 + \Omega_1)^{-1}\Omega_1\hat{\beta}.$$

This expression illustrates the compromise between information in the prior and the likelihood. The posterior mean B is a precision-weighted average of $\hat{\beta}$ and the prior mean of zero. The prior distribution stabilizes the required matrix inversion when $\mathbf{X}$ is rank deficient or nearly so, as often occurs in the presence of collinearity.

Finally prior distributions can be used to model relationships between complicated data structures. HIERARCHICAL MODELS, which are often to model nested data, provide a good example. In a hierarchical model, observations in a subgroup of the data set (e.g. student test scores for students in the same school) are modeled as conditionally independent given the prior for that subgroup. The prior parameters for the various subgroups are linked by a hyperprior distribution. The hyperprior allows subgroup parameters to "borrow strength" from other subgroups so that parameters of sparse subgroups can be accurately estimated. See the entry on HIERARCHICAL MODEL for details.

References

Efron, B. (1986). Why isn't everyone a Bayesian? (c/r: P5-11; p330-331). *The American Statistician* **40**, 1–5.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 1461–1473.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules (corr: 1998v93 p412). *Journal of the American Statistical Association* **91**, 1343–1370.

Table 1: Conjugate prior distributions for several common likelihoods.

Likelihood	Conjugate Prior
Univariate Normal	Normal (mean parameter)
	Gamma (inverse variance parameter)
Multivariate Normal	Multivariate Normal (mean vector)
	Wishart (inverse variance matrix)
Binomial	Beta
Poisson/Exponential	Gamma
Multinomial	Dirichlet