Principles of Bayesian Inference

Sudipto Banerjee¹ and Andrew O. Finley²

¹ Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
² Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

Bayesian principles

- The key to Bayesian inference is “learning” or “updating” of prior beliefs. Thus, posterior information ≈ prior information.

- Is the classical approach wrong? That may be a controversial statement, but it certainly is fair to say that the classical approach is limited in scope.

- The Bayesian approach expands the class of models and easily handles:
  - repeated measures
  - unbalanced or missing data
  - nonhomogenous variances
  - multivariate data
  - and many other settings that are precluded (or much more complicated) in classical settings.

Basics of Bayesian inference

- We start with a model (likelihood) \( f(\mathbf{y} | \theta) \) for the observed data \( \mathbf{y} = (y_1, \ldots, y_n) \) given unknown parameters \( \theta \) (perhaps a collection of several parameters).

- Add a prior distribution \( p(\theta | \lambda) \), where \( \lambda \) is a vector of hyper-parameters.

- The posterior distribution of \( \theta \) is given by:

\[
 p(\theta | \mathbf{y}, \lambda) = \frac{p(\theta | \lambda) \times f(\mathbf{y} | \theta)}{f(\mathbf{y} | \lambda)}
\]

We refer to this formula as Bayes Theorem.

Bayesian inference: point estimation

- Point estimation is easy: simply choose an appropriate distribution summary: posterior mean, median or mode.

- Mode sometimes easy to compute (no integration, simply optimization), but often misrepresents the “middle” of the distribution – especially for one-tailed distributions.

- Mean: easy to compute. It has the “opposite effect” of the mode – chases tails.

- Median: probably the best compromise in being robust to tail behaviour although it may be awkward to compute as it needs to solve:

\[
 \int_{-\infty}^{\theta_{median}} p(\theta | \mathbf{y})d\theta = \frac{1}{2}
\]
Sampling-based inference

We will compute the posterior distribution \( p(\theta \mid y) \) by drawing samples from it. This replaces numerical integration (quadrature) by “Monte Carlo integration”.

One important advantage: we only need to know \( p(\theta \mid y) \) up to the proportionality constant.

Suppose \( \theta = (\theta_1, \theta_2) \) and we know how to sample from the marginal posterior distribution \( p(\theta_2 \mid y) \) and the conditional distribution \( P(\theta_1 \mid \theta_2, y) \).

How do we draw samples from the joint distribution: \( p(\theta_1, \theta_2 \mid y) \)?

Example: Consider a single data point \( y \) from a Normal distribution: \( y \sim N(\theta, \sigma^2) \); assume \( \sigma \) is known.

\[
f(y \mid \theta) = N(y \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2\sigma^2}(y - \theta)^2 \right)
\]

\( \theta \sim N(\mu, \tau^2) \), i.e. \( p(\theta) = N(\theta \mid \mu, \tau^2) \); \( \mu, \tau^2 \) are known.

Posterior distribution of \( \theta \)

\[
p(\theta \mid y) \propto N(\theta \mid \mu, \tau^2) \times N(y \mid \theta, \sigma^2) = N\left( \theta \mid \frac{n}{\sigma^2 + \tau^2} \mu + \frac{\tau^2}{\sigma^2 + \tau^2} y, \frac{\sigma^2 + \tau^2}{\sigma^2 + \tau^2} \right).
\]

A simple example: Normal data and normal priors

Interpret: Posterior mean is a weighted mean of prior mean and data point.

Example: Let \( Y \) be the number of successes in \( n \) independent trials.

\( P(Y = y \mid \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n - y} \)

Prior: \( p(\theta) = Beta(\theta | a, b) \):

\[ p(\theta) \propto \theta^{a - 1} (1 - \theta)^{b - 1}. \]

Prior mean: \( \mu = a / (a + b) \); Variance \( ab / ((a + b)^2 (a + b + 1)) \)

Posterior distribution of \( \theta \)

\[
p(\theta \mid y) = Beta(\theta | a + y, b + n - y)
\]

Another simple example: The Beta-Binomial model

We do this in two stages using combination sampling:

- First draw \( \theta^{(j)}_2 \sim p(\theta_2 \mid y) \), \( j = 1, \ldots, M \).
- Next draw \( \theta^{(j)}_1 \sim p(\theta_1 \mid \theta^{(j)}_2, y) \).

This sampling scheme produces exact samples, \( \{\theta^{(j)}_1, \theta^{(j)}_2\}_{j=1}^M \) from the posterior distribution \( p(\theta_1, \theta_2 \mid y) \).

Gelfand and Smith (JASA, 1990) demonstrated automatic marginalization: \( \{\theta^{(j)}_1\}_{j=1}^M \) are samples from \( p(\theta_1 \mid y) \) and (of course) \( \{\theta^{(j)}_2\}_{j=1}^M \) are samples from \( p(\theta_2 \mid y) \).

In effect, composition sampling has performed the following “integration”:

\[
p(\theta_1 \mid y) = \int p(\theta_1 \mid \theta_2, y) p(\theta_2 \mid y) d\theta_2.
\]
Bayesian predictions

Suppose we want to predict new observations, say \( \tilde{y} \), based upon the observed data \( y \). We will specify a joint probability model \( p(\tilde{y}, y | \theta) \), which defines the conditional predictive distribution:

\[
p(\tilde{y} | y, \theta) = \frac{p(\tilde{y}, y | \theta)}{p(y | \theta)}.
\]

Bayesian predictions follow from the posterior predictive distribution that averages out the \( \theta \) from the conditional predictive distribution with respect to the posterior:

\[
p(\tilde{y} | y) = \int p(\tilde{y} | y, \theta)p(\theta | y)d\theta.
\]

This can be evaluated using composition sampling:

- First obtain: \( \theta^{(j)} \sim p(\theta | y), j = 1, \ldots, M \)
- For \( j = 1, \ldots, M \) sample \( \tilde{y}^{(j)} \sim p(\tilde{y} | y, \theta^{(j)}) \)
- The \( \{\tilde{y}^{(j)}\}_{j=1}^M \) are samples from the posterior predictive distribution \( p(\tilde{y} | y) \).

Some remarks on sampling-based inference

Direct Monte Carlo: Some algorithms (e.g., composition sampling) can generate independent samples exactly from the posterior distribution. In these situations there are NO convergence problems or issues. Sampling is called \textit{exact}.

Markov Chain Monte Carlo (MCMC): In general, exact sampling may not be possible/feasible. MCMC is a far more versatile set of algorithms that can be invoked to fit more general models. Note: anywhere where direct Monte Carlo applies, MCMC will provide excellent results too.

Convergence issues: There is no free lunch! The power of MCMC comes at a cost. The initial samples do not necessarily come from the desired posterior distribution. Rather, they need to converge to the true posterior distribution. Therefore, one needs to assess convergence, discard output before the convergence and retain only post-convergence samples. The time of convergence is called \textit{burn-in}.

Diagnosing convergence: Usually a few parallel chains are run from rather different starting points. The sample values are plotted (called \textit{trace-plots}) for each of the chains. The time for the chains to “mix” together is taken as the time for convergence.

Good news! All this is automated in WinBUGS. So, as users, we need to only configure how to specify good Bayesian models and implement them in WinBUGS.