Principles of Bayesian Inference

Alan Gelfand\textsuperscript{1} and Andrew O. Finley\textsuperscript{2}

\textsuperscript{1} Department of Statistical Science, Duke University, Durham, North Carolina.
\textsuperscript{2} Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan.

September 9, 2014

Basics of Bayesian inference

Calculations (numerical and algebraic) are usually required only up to a proportionally constant. We, therefore, write the posterior as:

\[ p(\theta | y, \lambda) \propto p(\theta | \lambda) \times f(y | \theta). \]

If \( \lambda \) are known/fixed, then the above represents the desired posterior. If, however, \( \lambda \) are unknown, we assign a prior, \( p(\lambda) \), and seek:

\[ p(\theta, \lambda | y) \propto p(\theta | \lambda) p(\lambda) f(y | \theta). \]

The proportionality constant does not depend upon \( \theta \) or \( \lambda \):

\[ \frac{1}{p(y)} = \int p(\lambda)p(\theta | \lambda)f(y | \theta)d\lambda d\theta. \]

The above represents a joint posterior from a hierarchical model. The marginal posterior distribution for \( \theta \) is:

\[ p(\theta | y) = \int p(\lambda)p(\theta | \lambda)f(y | \theta)d\lambda. \]

Bayesian principles

- Classical statistics: model parameters are fixed and unknown.
- A Bayesian thinks of parameters as random, and thus having distributions (just like the data).
- A Bayesian writes down a prior guess for parameter(s) \( \theta \), say \( p(\theta) \). They then combines this with the information provided by the observed data \( y \) to obtain the posterior distribution of \( \theta \), which we denote by \( p(\theta | y) \).
- All statistical inferences (point and interval estimates, hypothesis tests) then follow from posterior summaries. For example, the posterior means/medians/modes offer point estimates of \( \theta \), while the quantiles yield credible intervals.

We start with a model (likelihood) \( f(y | \theta) \) for the observed data \( y = (y_1, \ldots, y_n)' \) 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:

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

We refer to this formula as Bayes Theorem.

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 prior distribution – especially for one-tailed distributions.

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

\[ \int_{-\infty}^{\theta_{median}} p(\theta | y) d\theta = \frac{1}{2}. \]
Bayesian inference: interval estimation

- The most popular method of inference in practical Bayesian modelling is interval estimation using credible sets. A 100(1−α)% credible set C for θ is a set that satisfies:

\[ P(θ ∈ C | y) = \int_{C} p(θ | y) dθ ≥ 1 − α. \]

- The most popular credible set is the simple equal-tail interval estimate (qL, qU) such that:

\[ \int_{qL}^{qU} p(θ | y) dθ = \frac{α}{2} = \int_{qL}^{qU} p(θ | y) dθ \]

Then clearly \( P(θ ∈ (qL, qU) | y) = 1 − α. \)

- This interval is relatively easy to compute and has a direct interpretation: The probability that θ lies between (qL, qU) is 1 − α. The frequentist interpretation is extremely convoluted.

Another simple example: Normal data and normal priors

Example: Consider a single data point \( y \) from a Normal distribution: \( y ∼ N(μ, τ^2) \). Assume \( τ^2 \) is known.

The direct estimate is shrunk towards the prior.

\[ p(y|θ) = N(θ| y, σ^2) = \frac{1}{σ\sqrt{2π}} \exp\left(-\frac{(y−θ)^2}{2σ^2}\right) \]

Now set the prior for \( θ ∼ N(μ, τ^2) \), i.e. \( p(θ) = N(θ| μ, τ^2) \);

\[ µ, τ^2 \] are known.

Posterior distribution of \( θ \)

\[ p(θ | y) ∝ N(θ| μ, τ^2) \times N(y | θ, σ^2) = N\left(θ | \frac{1}{τ^2 + σ^2} μ + \frac{1}{τ^2 + σ^2} y, \frac{1}{τ^2 + σ^2}\right) \]

Prior: \( p(θ) = Beta(θ| a, b) \):

\[ p(θ) ∝ θ^{a-1}(1−θ)^{b−1}. \]

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

Posterior distribution of \( θ \)

\[ p(θ | y) = Beta(θ| a+y, b+n−y) \]

Sampling-based inference

- In practice, we will compute the posterior distribution \( p(θ | y) \) by drawing samples from it. This replaces numerical integration (quadrature) by “Monte Carlo integration”.

- One important advantage: we only need to know \( p(θ | y) \) up to the proportionality constant.

- Suppose \( θ = (θ_1, θ_2) \) and we know how to sample from the marginal posterior distribution \( p(θ_2 | y) \) and the conditional distribution \( P(θ_1 | θ_2, y) \).

- How do we draw samples from the joint distribution: \( p(θ_1, θ_2 | y) \)?

We do this in two stages using composition sampling:

- First draw \( θ_2^{(j)} \) \( p(θ_2 | y) \), \( j = 1, ..., M \).

- Next draw \( θ_1^{(j)} \) \( p\left(θ_1 | θ_2^{(j)}, y\right) \).

This sampling scheme produces exact samples, \( \{θ_1^{(1)}, θ_2^{(1)}\}_{j=1}^{M} \) from the posterior distribution \( p(θ_1, θ_2 | y) \).

Gelfand and Smith (JASA, 1990) demonstrated automatic marginalization: \( \{θ_2^{(1)}\}_{j=1}^{M} \) are samples from \( p(θ_1 | y) \) and (of course!) \( \{θ_1^{(1)}\}_{j=1}^{M} \) are samples from \( p(θ_2 | y) \).

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

\[ p(θ_1 | y) = \int p(θ_1 | θ_2, y) p(θ_2 | y) dθ_2. \]
Bayesian predictions

- Suppose we want to predict new observations, say $\tilde{y}$, based upon the observed data $y$.
- 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 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 burn-in.

- Diagnosing convergence: Usually a few parallel chains are run from rather different starting points. The sample values are plotted (called trace-plots) for each of the chains. The time for the chains to “mix” together is taken as the time for convergence.
- Good news! Many modeling frameworks are automated in freely available software. So, as users, we need to only configure how to specify good Bayesian models and

Find a wide variety of R packages dealing with Bayesian inference here:

http://cran.r-project.org/web/views/Bayesian.html

Here’s a nice rant on “Why I love R” http://www.sr.bham.ac.uk/~ajrs/talks/why_I_love_R.pdf