Principles of Bayesian Inference

Sudipto Banerjee[1] and Andrew O. Finley[2]

[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

- Classical statistics: model parameters are *fixed* and *unknown*.
- A Bayesian thinks of parameters as random, and thus having distributions (just like the data). We can thus think about unknowns for which no reliable frequentist experiment exists, e.g. $\theta =$ proportion of US men with untreated prostate cancer.
- A Bayesian writes down a *prior* guess for parameter(s) $\boldsymbol{\theta}$, say $p(\boldsymbol{\theta})$. He then combines this with the information provided by the observed data $\mathbf{y}$ to obtain the *posterior* distribution of $\boldsymbol{\theta}$, which we denote by $p(\boldsymbol{\theta} \mid \mathbf{y})$.
- 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 $\boldsymbol{\theta}$, while the quantiles yield credible intervals.

- The key to Bayesian inference is "learning" or "updating" of prior beliefs. Thus, posterior information $\geq$ 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.

- We start with a model (likelihood) $f(\mathbf{y} \mid \boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)'$ given unknown parameters $\boldsymbol{\theta}$ (perhaps a collection of several parameters).
- Add a prior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of hyper-parameters.
- The posterior distribution of $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda}) = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times f(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y} \mid \boldsymbol{\lambda})} = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times f(\mathbf{y} \mid \boldsymbol{\theta})}{\int f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) d\boldsymbol{\theta}}.$$

  We refer to this formula as *Bayes Theorem*.

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

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda}) \propto p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times f(\mathbf{y} \mid \boldsymbol{\theta}).$$

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

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \mathbf{y}) \propto p(\boldsymbol{\lambda}) p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) f(\mathbf{y} \mid \boldsymbol{\theta}).$$

  The proportionality constant does not depend upon $\boldsymbol{\theta}$ or $\boldsymbol{\lambda}$:

$$\frac{1}{p(\mathbf{y})} = \frac{1}{\int p(\boldsymbol{\lambda}) p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) f(\mathbf{y} \mid \boldsymbol{\theta}) d\boldsymbol{\lambda} d\boldsymbol{\theta}}$$

- The above represents a *joint* posterior from a *hierarchical model*. The *marginal* posterior distribution for $\boldsymbol{\theta}$ is:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \int p(\boldsymbol{\lambda}) p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) f(\mathbf{y} \mid \boldsymbol{\theta}) d\boldsymbol{\lambda}.$$

- 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:

$$\int_{-\infty}^{\boldsymbol{\theta}_{median}} p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} = \frac{1}{2}.$$

- The most popular method of inference in practical Bayesian modelling is interval estimation using *credible sets*. A $100(1-\alpha)\%$ credible set C for $\boldsymbol{\theta}$ is a set that satisfies:

$$P(\boldsymbol{\theta} \in C \,|\, \mathbf{y}) = \int_C p(\boldsymbol{\theta} \,|\, \mathbf{y}) d\boldsymbol{\theta} \geq 1-\alpha.$$

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

$$\int_{-\infty}^{q_L} p(\theta \,|\, \mathbf{y}) d\theta = \frac{\alpha}{2} = \int_{q_U}^{\infty} p(\theta \,|\, \mathbf{y}) d\theta$$

  Then clearly $P(\theta \in (q_L, q_U) \,|\, \mathbf{y}) = 1-\alpha$.
- This interval is relatively easy to compute and has a direct interpretation: The probability that θ lies between (q_L, q_U) is $1-\alpha$. The frequentist interpretation is extremely convoluted.

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

$$f(y|\theta) = N(y \,|\, \theta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(y-\theta)^2)$$

- $\theta \sim N(\mu, \tau^2)$, i.e. $p(\theta) = N(\theta \,|\, \mu, \tau^2)$; μ, τ^2 are known.
- Posterior distribution of θ

$$\begin{aligned} p(\theta|y) &\propto N(\theta \,|\, \mu, \tau^2) \times N(y \,|\, \theta, \sigma^2) \\ &= N\left(\theta \,|\, \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\mu + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}y, \frac{1}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\right) \\ &= N\left(\theta \,|\, \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right). \end{aligned}$$

- Interpret: Posterior mean is a weighted mean of prior mean and data point.
- The direct estimate is shrunk towards the prior.
- What if you had n observations instead of one in the earlier set up? Say $\mathbf{y} = (y_1, \ldots, y_n)'$, where $y_i \overset{iid}{\sim} N(0, \sigma^2)$.
- $\bar{y}$ is a *sufficient statistic* for μ; $\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- Posterior distribution of θ

$$\begin{aligned} p(\theta \,|\, \mathbf{y}) &\propto N(\theta \,|\, \mu, \tau^2) \times N\left(\bar{y} \,|\, \theta, \frac{\sigma^2}{n}\right) \\ &= N\left(\theta \,|\, \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\mu + \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\bar{y}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right) \\ &= N\left(\theta \,|\, \frac{\sigma^2}{\sigma^2+n\tau^2}\mu + \frac{n\tau^2}{\sigma^2+n\tau^2}\bar{y}, \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right) \end{aligned}$$

- Example: Let Y be the number of successes in n independent trials.

$$P(Y=y|\theta) = f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- 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 θ

$$p(\theta|y) = Beta(\theta|a+y, b+n-y)$$

- We will compute the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ by drawing samples from it. This replaces numerical integration (quadrature) by "Monte Carlo integration".
- One important advantage: we only need to know $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ up to the proportionality constant.
- Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and we know how to sample from the *marginal posterior distribution* $p(\boldsymbol{\theta}_2|\mathbf{y})$ and the *conditional distribution* $P(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2, \mathbf{y})$.
- How do we draw samples from the joint distribution: $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \,|\, \mathbf{y})$?

- We do this in two stages using *composition sampling*:
  - First draw $\boldsymbol{\theta}_2^{(j)} \sim p(\boldsymbol{\theta}_2 \,|\, \mathbf{y}),\ j = 1, \ldots M.$
  - Next draw $\boldsymbol{\theta}_1^{(j)} \sim p\left(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2^{(j)}, \mathbf{y}\right)$.
- This sampling scheme produces *exact* samples, $\{\boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_2^{(j)}\}_{j=1}^M$ from the posterior distribution $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \,|\, \mathbf{y})$.
- Gelfand and Smith (*JASA*, 1990) demonstrated *automatic marginalization*: $\{\boldsymbol{\theta}_1^{(j)}\}_{j=1}^M$ are samples from $p(\boldsymbol{\theta}_1 \,|\, \mathbf{y})$ and (of course!) $\{\boldsymbol{\theta}_2^{(j)}\}_{j=1}^M$ are samples from $p(\boldsymbol{\theta}_2 \,|\, \mathbf{y})$.
- In effect, composition sampling has performed the following "integration":

$$p(\boldsymbol{\theta}_1 \,|\, \mathbf{y}) = \int p(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2, \mathbf{y}) p(\boldsymbol{\theta}_2 \,|\, \mathbf{y}) d\boldsymbol{\theta}.$$

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

$$p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\tilde{\mathbf{y}}, \mathbf{y} \,|\, , \boldsymbol{\theta})}{p(\mathbf{y} \,|\, \boldsymbol{\theta})}.$$

- Bayesian predictions follow from the *posterior predictive* distribution that averages out the $\boldsymbol{\theta}$ from the conditional predictive distribution with respect to the posterior:

$$p(\tilde{\mathbf{y}} \,|\, \mathbf{y}) = \int p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) d\boldsymbol{\theta}.$$

- This can be evaluated using composition sampling:
  - First obtain: $\boldsymbol{\theta}^{(j)} \sim p(\boldsymbol{\theta} \,|\, \mathbf{y}),\ j = 1, \ldots M$
  - For $j = 1, \ldots, M$ sample $\tilde{\mathbf{y}}^{(j)} \sim p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(j)})$
- The $\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^{M}$ are samples from the posterior predictive distribution $p(\tilde{\mathbf{y}} \,|\, \mathbf{y})$.

- 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! 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.