

Principles of Bayesian Inference

Sudipto Banerjee

University of Minnesota

July 20th, 2008

- 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. θ = proportion of US men with untreated prostate cancer.
- A Bayesian writes down a *prior* guess for parameter(s) θ , say $p(\theta)$. He then combines this with the information provided by the observed data \mathbf{y} to obtain the *posterior* distribution of θ , which we denote by $p(\theta | \mathbf{y})$.
- 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 θ , 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) $f(\mathbf{y} | \boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \dots, y_n)'$ given unknown parameters $\boldsymbol{\theta}$ (perhaps a collection of several parameters).
- Add a prior distribution $p(\boldsymbol{\theta} | \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of hyper-parameters.
- The posterior distribution of $\boldsymbol{\theta}$ is given by:

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

We refer to this formula as *Bayes Theorem*.

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

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

- 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:

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

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

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

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

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

- 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:

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

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

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

- 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*.

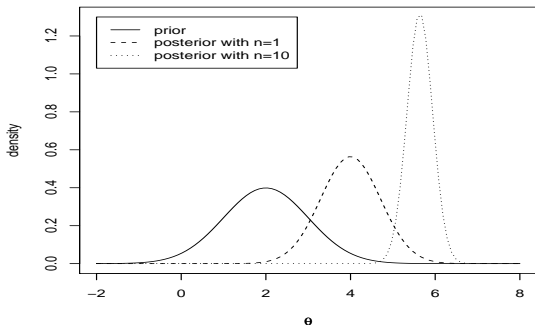
$$f(y|\theta) = N(y|\theta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - \theta)^2\right)$$

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

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

- 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 $\mathbf{y} = (y_1, \dots, y_n)'$, where $y_i \stackrel{iid}{\sim} N(\theta, \sigma^2)$.
- \bar{y} is a *sufficient statistic* for μ ; $\bar{y} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$
- Posterior distribution of θ

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



- Example: $\mu = 2$, $\bar{y} = 6$, $\tau^2 = \sigma^2 = 1$.
- When $n = 1$, the prior and the likelihood receive equal weight, so the posterior mean is $4 = \frac{2+6}{2}$
- When $n = 10$, the data dominate the prior and the posterior mean approaches \bar{y} .
- The posterior variance also shrinks as n gets larger; “collapses” to \bar{y} .

- Consider the problem of estimating the current weight of a group of people. A sample of 10 people were taken and their average weight was calculated as $\bar{y} = 176$ lbs. Assume that the population standard deviation was known as $\sigma = 3$. Assuming that the data y_1, \dots, y_{10} came from a $N(\theta, \sigma^2)$ population perform the following:
- Obtain a 95% confidence interval for θ using classical methods.
- Assume a prior distribution for θ of the form $N(\mu, \tau^2)$. Obtain 95% posterior credible intervals for θ for each of the cases: (a) $\mu = 176, \tau = 8$; (b) $\mu = 176, \tau = 1000$ (c) $\mu = 0, \tau = 1000$. Which case gives results closest to that obtained in the classical method?

- 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: $p(\theta) = \text{Beta}(\theta|a, b)$:

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

- Prior mean: $\mu = a/(a + b)$; Variance $ab/((a + b)^2(a + b + 1))$
- Posterior distribution of θ

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

- 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 $\theta_2^{(j)} \sim p(\theta_2 | \mathbf{y})$, $j = 1, \dots, M$.
 - Next draw $\theta_1^{(j)} \sim p(\theta_1 | \theta_2^{(j)}, \mathbf{y})$.
- This sampling scheme produces *exact* samples, $\{\theta_1^{(j)}, \theta_2^{(j)}\}_{j=1}^M$ from the posterior distribution $p(\theta_1, \theta_2 | \mathbf{y})$.
- Gelfand and Smith (*JASA*, 1990) demonstrated *automatic marginalization*: $\{\theta_1^{(j)}\}_{j=1}^M$ are samples from $p(\theta_1 | \mathbf{y})$ and (of course!) $\{\theta_2^{(j)}\}_{j=1}^M$ are samples from $p(\theta_2 | \mathbf{y})$.
- In effect, composition sampling has performed the following “integration”:

$$p(\theta_1 | \mathbf{y}) = \int p(\theta_1 | \theta_2, \mathbf{y})p(\theta_2 | \mathbf{y})d\theta.$$

- 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 $p(\tilde{\mathbf{y}}, \mathbf{y} | \theta)$, which defines the *conditional predictive distribution*:

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

- Bayesian predictions follow from the *posterior predictive* distribution that averages out the θ from the conditional predictive distribution with respect to the posterior:

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

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

- Suppose that $\theta = (\theta_1, \theta_2)$ and we seek the posterior distribution $p(\theta_1, \theta_2 | \mathbf{y})$.
- For many interesting hierarchical models, we have access to *full conditional distributions* $p(\theta_1 | \theta_2, \mathbf{y})$ and $p(\theta_2 | \theta_1, \mathbf{y})$.
- The *Gibbs sampler* proposes the following sampling scheme. Set starting values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$ For $j = 1, \dots, M$
 - Draw $\theta_1^{(j)} \sim p(\theta_1 | \theta_2^{(j-1)}, \mathbf{y})$
 - Draw $\theta_2^{(j)} \sim p(\theta_2 | \theta_1^{(j)}, \mathbf{y})$
- This constructs a *Markov Chain* and, after an initial “burn-in” period when the chains are trying to find their way, the above algorithm guarantees that $\{\theta_1^{(j)}, \theta_2^{(j)}\}_{j=M_0+1}^M$ will be samples from $p(\theta_1, \theta_2 | \mathbf{y})$, where M_0 is the burn-in period..

- More generally, if $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$ are the parameters in our model, we provide a set of initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_p^{(0)})$ and then performs the j -th iteration, say for $j = 1, \dots, M$, by updating successively from the *full conditional* distributions:

$$\boldsymbol{\theta}_1^{(j)} \sim p(\boldsymbol{\theta}_1^{(j)} \mid \boldsymbol{\theta}_2^{(j-1)}, \dots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

$$\boldsymbol{\theta}_2^{(j)} \sim p(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_3^{(j)}, \dots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

...

(the generic k^{th} element)

$$\boldsymbol{\theta}_k^{(j)} \sim p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_{k-1}^{(j)}, \boldsymbol{\theta}_{k+1}^{(j)}, \dots, \boldsymbol{\theta}_p^{(j-1)}, \mathbf{y})$$

...

$$\boldsymbol{\theta}_p^{(j)} \sim p(\boldsymbol{\theta}_p \mid \boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_{p-1}^{(j)}, \mathbf{y})$$

- In principle, the Gibbs sampler will work for extremely complex hierarchical models. The only issue is sampling from the full conditionals. They may not be amenable to easy sampling – when these are not in closed form. A more general and extremely powerful - and often easier to code - algorithm is the Metropolis-Hastings (MH) algorithm.
- This algorithm also constructs a Markov Chain, but does not necessarily care about full conditionals.

- The Metropolis-Hastings algorithm: Start with a initial value for $\theta = \theta^{(0)}$. Select a *candidate* or *proposal* distribution from which to propose a value of θ at the j -th iteration: $\theta^{(j)} \sim q(\theta^{(j-1)}, \nu)$. For example, $q(\theta^{(j-1)}, \nu) = N(\theta^{(j-1)}, \nu)$ with ν fixed.

- Compute

$$r = \frac{p(\theta^* | \mathbf{y})q(\theta^{(j-1)} | \theta^*, \nu)}{p(\theta^{(j-1)} | \mathbf{y})q(\theta^* | \theta^{(j-1)}, \nu)}$$

- If $r \geq 1$ then set $\theta^{(j)} = \theta^*$. If $r \leq 1$ then draw $U \sim (0, 1)$. If $U \leq r$ then $\theta^{(j)} = \theta^*$. Otherwise, $\theta^{(j)} = \theta^{(j-1)}$.
- Repeat for $j = 1, \dots, M$. This yields $\theta^{(1)}, \dots, \theta^{(M)}$, which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio r of the sampler through the iterations. Rough recommendations: for vector updates $r \approx 20\%$, for scalar updates $r \approx 40\%$. This can be controlled by “tuning” ν .
- Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.

- 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.