2021 Advanced Regression Methods for Independent Data

BIOSTAT/STAT 570

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

CHAPTER 3: BAYESIAN INFERENCE

Outline

Motivation and Posterior Summarization

Prior Choice

Asymptotic Behavior

Computation

- Conjugate Analysis
- Laplace Approximations
- Direct Sampling
- Gaussian Quadrature
- Importance Sampling
- INLA
- MCMC

Bayes Factors

Motivation and Posterior Summarization

Bayesian Inference

In the Bayesian approach to inference all unknown quantities contained in a probability model for the observed data are treated as random variables.

These unknowns may include:

- missing data,
- the true covariate value in an errors-in-variables setting,
- the failure time of a censored survival observation,
- predictions for new observations.

Bayesian Inference

Inference is made through the **posterior** probability distribution of $\boldsymbol{\theta}$ after observing $\boldsymbol{y}$, and is determined from **Bayes theorem:**

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

where, for continuous $\boldsymbol{\theta}$, the normalizing constant is

$$p(\boldsymbol{y}) = \int_{\theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

This is the marginal probability of the observed data averaged over the assumed model, i.e., the likelihood and prior – before we saw data, would we **really believe** this was the probability of observing $\boldsymbol{y}$?

Bayesian inferential probabilities are for random $\boldsymbol{\theta}$ conditional on the data, which is in stark contrast to frequentist inferential probabilities which are over hypothetical sampling of new datasets for fixed $\boldsymbol{\theta}$.

Bayesian Inference

Ignoring the constant gives

$$\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{y})}_{\text{Posterior}} \propto \underbrace{p(\boldsymbol{y} \mid \boldsymbol{\theta})}_{\text{Likelihoood}} \times \underbrace{\pi(\boldsymbol{\theta})}_{\text{Prior}}$$

The use of the posterior distribution for inference is very intuitively appealing since it probabilistically combines information on the parameters arising from the data and from prior beliefs.

An important observation is that,

$$\pi(\boldsymbol{\theta}) = 0 \qquad \Rightarrow \qquad p(\boldsymbol{\theta} \mid \boldsymbol{y}) = 0$$

regardless of any realization of the observed data.

This has important consequences for prior specification and clearly shows that great care should be taken in excluding parts of the parameter space a priori – you never get them back!

Less obviously, when we pick particular likelihoods, we are constraining the data in some way (think moment assumptions, for example) – it gets scary quite quickly!

Sequential Updating

Suppose first that $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ represent the current totality of data; the posterior is given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1, \boldsymbol{y}_2)}. \tag{1}$$

Now suppose that we are at a previous time point at which only $\boldsymbol{y}_1$ are available, the posterior in this case is

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1) = \frac{p(\boldsymbol{y}_1 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1)}.$$

When $\boldsymbol{y}_2$ becomes available, the “prior” for these data corresponds to $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$ since it represents the current beliefs concerning $\boldsymbol{\theta}$.

Then update via

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_1)}{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1)}. \tag{2}$$

Identical inference in each case.

Inferential Summaries

To summarizes the typically multivariate posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, marginal distributions for parameters of interest may be considered.

For example the univariate marginal distribution for a component θ_i is,

$$p(\theta_i \mid \boldsymbol{y}) = \int_{\theta_{-i}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{\theta}_{-i}, \tag{3}$$

where $\boldsymbol{\theta}_{-i}$ is the vector $\boldsymbol{\theta}$ excluding θ_i.

Posterior moments may be evaluated from the marginal distributions; for example the posterior mean is,

$$\mathrm{E}[\theta_i \mid \boldsymbol{y}] = \int_{\theta_i} \theta_i p(\theta_i \mid \boldsymbol{y}) \, d\theta_i. \tag{4}$$

In frequentist inference, we have a point estimate and a measure of uncertainty (both reflections of the sampling distribution), in Bayesian inference, the fundamental summary is the posterior distribution, then we choose how to summarize.

Inferential Summaries

Further summarization may be carried out to yield the 100×q% posterior quantile, $\theta_i(q)$ $(0 < q < 1)$ by solving

$$\int_{-\infty}^{\theta_i(q)} p(\theta_i \mid \boldsymbol{y}) \, d\theta_i. \tag{5}$$

In particular, the posterior median, $\theta_i(0.5)$, will often provide an adequate summary of the location of the posterior marginal distribution.

A 100×p% equi-tailed credible interval $(0 < p < 1)$ is provided by

$$[\ \theta_i\{(1-p)/2\}, \theta_i\{(1+p)/2\}\ ].$$

Such intervals are usually reported though in some cases it which the posterior is skewed one may wish to instead calculate a highest posterior density (HPD) interval in which points inside the interval have higher posterior density than those outside the interval (such an interval is also the shortest credible interval).

Predictive Distributions

Another useful inferential quantity is the **predictive** distributions for future observations ***z*** which is given, under conditional independence, by

$$\begin{aligned} p(\boldsymbol{z} \mid \boldsymbol{y}) &= \int_\theta p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \\ &= \int_\theta p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \\ &= \int_\theta \underbrace{p(\boldsymbol{z} \mid \boldsymbol{\theta})}_{\text{From conditional independence}} \quad p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \qquad (6) \\ &= \mathrm{E}_{\theta|y}\left[p(\boldsymbol{z} \mid \boldsymbol{\theta})\right] \end{aligned}$$

This shows the advantage of the use of probability under a Bayesian approach – less obvious to proceed under likelihood – $p(\boldsymbol{z} \mid \widehat{\boldsymbol{\theta}})$ is the obvious candidate, but how to capture parameter uncertainty?

This clearly assumes that the system under study is “stable” so that the likelihood for future observations is still the relevant data generation mechanism.

Bayesian Inference – In Practice

Bayesian inference is deceptively simple to describe probabilistically, but there have been two major obstacles to its routine use:

- The first is how to specify prior distributions.
- The second is how to do the computation evaluate the integrals required for inference, for example, (3)–(6), given that for most models, these are analytically intractable.

In addition, when compared to inference based on quasi-likelihood or sandwich estimation, we need to specify a full probability model for the data, i.e., a likelihood specification.

Normal Example: Estimation

Suppose we have

$$Y_i|\theta \sim_{iid} \mathrm{N}(\theta, \sigma^2), \quad i = 1, \dots, n,$$

with σ^2 assumed known and θ unknown.

Recall that the MLE has sampling distribution,

$$\overline{Y} \mid \theta \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right).$$

Suppose the prior distribution is $\theta \sim \mathrm{N}(m, v)$, with m and v are known; the posterior distribution is

$$\theta \mid \boldsymbol{y} \sim \mathrm{N}\left(\overline{y} \times w + m \times (1 - w), \frac{\sigma^2}{n} \times w\right),$$

where

$$w = \frac{v}{v + \sigma^2/n} = \frac{nv}{nv + \sigma^2} = \frac{n}{n + \sigma^2 v^{-1}}.$$

Normal Example: Estimation

Think about cases:

- If $n = 0$ we recover the prior.
- If $v = 0$ the posterior is a point mass at the prior mean m.
- If $v^{-1} = 0$ we have an improper prior (more on these later), and frequentist and Bayesian estimates coincide.
- As $n \to \infty$ then $w \to 1$ (unless $v = 0$), and we hone in on correspondence between Bayes and MLE.

Normal Example: Estimation

One useful way of specifying the prior is as

$$\theta \sim \mathrm{N}\left(m, \frac{\sigma^2}{k}\right),$$

in which case k may be regarded as a prior sample size.

It is 'as if' we carried out an experiment with k observations and we observed a mean of m.

This gives $w = n/(n+k)$ and

$$\theta \mid \boldsymbol{y} \sim \mathrm{N}\left(\bar{y} \times \frac{n}{n+k} + m \times \frac{k}{n+k}, \frac{\sigma^2}{n} \times \frac{n}{n+k}\right).$$

Normal Example: Estimation

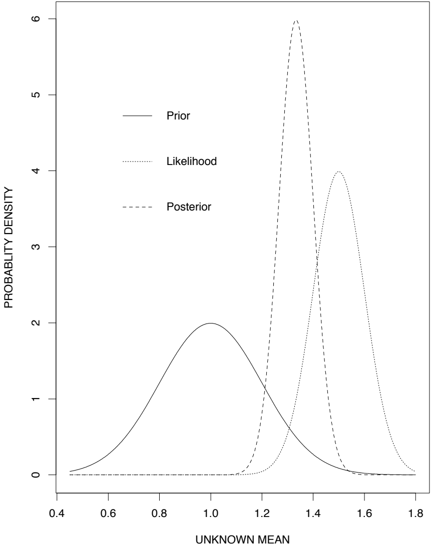


Figure 1: Normal likelihood ($\overline{y}$=1.5,n=10,σ=1), normal prior (m=1, k=5) and the resultant normal posterior.

Normal Example: Prediction

Suppose we wish to obtain the predictive density for a new random variable $Z \sim \mathrm{N}(\theta, \sigma^2)$.

Then

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

Not too tricky to show that

$$z \mid \boldsymbol{y} \sim \mathrm{N}\left(\mathrm{E}[\theta|\boldsymbol{y}], \sigma^2 + \mathrm{var}(\theta|\boldsymbol{y}) \right),$$

so that:

- the mean of the predictive distribution is the posterior mean and
- the variance is given by the sum of the measurement error σ^2 (we can't escape this) and the posterior uncertainty in θ.

Binomial Example: Estimation

- With the likelihood $Y|\theta \sim \text{Binomial}(N, \theta)$, the prior $\theta \sim \text{Beta}(a, b)$ is convenient.
- The posterior is

$$\begin{aligned} p(\theta|y) &\propto \Pr(y|\theta) \times p(\theta) \\ &\propto \theta^{y}(1-\theta)^{N-y} \times \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{N-y+b-1} \end{aligned}$$

  where we only keep track of the θ terms.
- We recognize this as the business end of a

$$\text{Beta}(y+a, N-y+b)$$

  distribution.
- We know what the normalizing constant must be, because we have a function which must integrate to 1.
- If $X \sim \text{Beta}(\alpha, \beta)$, the mean and variance are

$$\text{E}[X] = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \text{var}(X) = \frac{\text{E}[X](1-\text{E}[X])}{\alpha+\beta+1}.$$

The beta posterior

- Posterior mean is the weighted sum of the MLE and the prior mean:

$$\mathsf{E}[\theta|y] = \frac{y+a}{N+a+b} = \frac{y}{N}\frac{N}{N+a+b} + \frac{a}{a+b}\frac{a+b}{N+a+b}.$$

- We will rarely want to report a point estimate alone, whether it be a posterior mean or posterior median.

- Interval estimates are obtained in the obvious way.

- A simple way of performing testing of particular parameter values of interest is via examination of interval estimates.

- For example, does a 95% interval contain the value $\theta_0 = 0.5$?

Other Posterior Summaries

- In our beta-binomial example, a 90% posterior credible interval (θ_L, θ_U) results from the points

$$
\begin{aligned}
0.05 &= \int_0^{\theta_L} p(\theta|y)\, d\theta \\
0.95 &= \int_0^{\theta_U} p(\theta|y)\, d\theta
\end{aligned}
$$

- The quantiles of a beta are not available in closed form, but easy to evaluate in R:

```
y <- 7; N <- 10; a <- b <- 1
qbeta(c(0.05,0.5,0.95),y+a,N-y+b)
[1] 0.4356258 0.6761955 0.8649245
```

- The posterior median is 0.68 and a 90% credible interval is [0.44,0.86].

- The MLE is 0.70 and an asymptotic 90% confidence interval is $0.70 \pm 1.645 \times \sqrt{0.7 \times 0.3/10} = [0.46, 0.94]$.

Binomial Example: Prediction

The predictive distribution for a new trial in which $z = 0, 1, \dots, m$ denotes the number of successes and m the number of trials, is

$$\begin{aligned}
\Pr(z|y) &= \int_0^1 \Pr(z|\theta) \times p(\theta|y) d\theta \\
&= \int_0^1 \left(\begin{array}{c} M \\ z \end{array} \right) \theta^z (1-\theta)^{M-z} \\
\times & \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \theta^{y+a-1} (1-\theta)^{N-y+b-1} d\theta \\
&= \left(\begin{array}{c} M \\ z \end{array} \right) \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \int_0^1 \theta^{y+a+z-1} (1-\theta)^{N-y+b+M-z-1} d\theta \\
&= \left(\begin{array}{c} M \\ z \end{array} \right) \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \frac{\Gamma(a+y+z)\Gamma(b+N-y+M-z)}{\Gamma(a+b+N+M)}
\end{aligned}$$

for $z = 0, 1, \dots, M$.

Binomial Example: Prediction

A likelihood approach would take the predictive distribution as Binomial$(M, \widehat{\theta})$ with $\widehat{\theta} = y/N$ but this does not account for estimation uncertainty.

In general, we have sampling uncertainty (which we can't get away from) and estimation uncertainty.

We have just shown,

$$p(z|y) \quad = \quad \left(\begin{array}{c} M \\ z \end{array} \right) \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)} \frac{\Gamma(a+b+z)\Gamma(b+n-y+M-z)}{\Gamma(a+b+n+M)},$$

for $z = 0, \dots, M$, which is known as the beta-binomial distribution, an overdispersed binomial.

Binomial Example: Prediction

We have:

$$\begin{aligned} \mathsf{E}[Z|y] &= \mathsf{E}_{\theta|y}[\mathsf{E}[Z|\theta]] = \mathsf{E}_{\theta|y}[m \times \theta] \\ &= m \times \mathsf{E}[\theta|y] = m \times \frac{a+y}{a+b+n}, \end{aligned}$$

and

$$\begin{aligned} \text{var}(Z|y) &= \text{var}_{\theta|y}(\mathsf{E}[Z|\theta]) + \mathsf{E}_{\theta|y}[\text{var}(Z|\theta)] \\ &= m \times \mathsf{E}[\theta|y](1 - \mathsf{E}[\theta|y])\frac{a+b+n+m}{a+b+n+1}, \end{aligned}$$

showing the overdispersion (excess-binomial variation).

Note: This is the predictive distribution arising from a binomial likelihood and beta prior – the distribution may also be specified as a likelihood, parameterized in terms of the success probability and an overdispersion parameter, see Liang and McCullagh (1993).

Binomial Example: Prediction

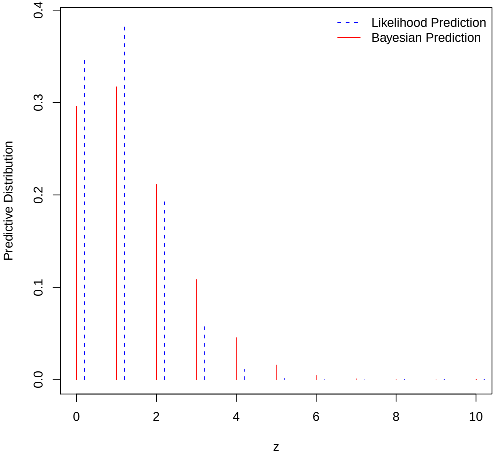


Figure 2: Likelihood and Bayesian predictive distribution of seeing $z = 0, 1, \ldots, M = 10$ successes, after observing $y = 2$ out of $N = 20$ successes (with $a = b = 1$).

Binomial Example: Prediction

The posterior and sampling distributions won't usually combine so conveniently.

In general, we may form a Monte Carlo estimate of the predictive distribution:

$$\begin{aligned} p(z|y) &= \int p(z|\theta)p(\theta|y)d\theta \\ &= \mathrm{E}_{\theta|y}[p(z|\theta)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} p(z|\theta^{(s)}) \end{aligned}$$

where $\theta^{(s)} \sim p(\theta|y)$, $s = 1, \ldots, S$, is a sample from the posterior.

This provides an estimate of the predictive distribution at the point z.

Binomial Example: Prediction

Alternatively, we may sample from $p(z|\theta^{(s)})$ a large number of times to reconstruct the predictive distribution:

$$\begin{aligned} \theta^{(s)}|y &\sim p(\theta|y),\ s=1,\dots,S && \text{Sample from posterior} \\ z^{(s)}|\theta^{(s)} &\sim p(z|\theta^{(s)}),\ s=1,\dots,S && \text{Sample from predictive} \end{aligned}$$

To give a sample $z^{(s)}$ from the posterior, this is illustrated in Figure 3.

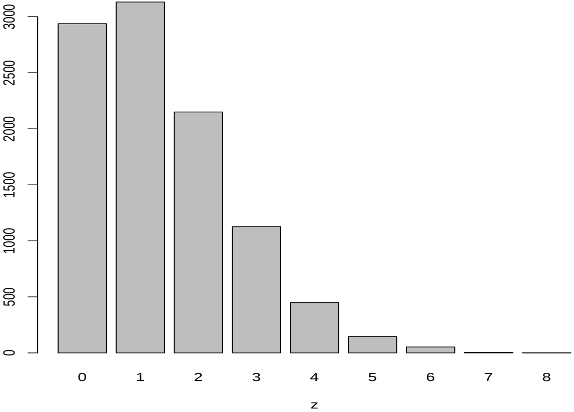


FIGURE 3: Sampling version of prediction in Figure 2, based on $S = 10,000$ samples.

Simple Examples

The above normal and binomial derivations are examples of conjugate Bayesian analyses in which the prior is in the same family as the posterior, unfortunately for most models such computationally convenient analyses are not possible.

However, conjugate calculations can be useful:

- For pedagogy and insight.
- As parts of other Bayesian computations, e.g., when doing Gibbs sampling.

Prior Choice

Prior Choice

We distinguish between two prior specification situations. In the first, which we label as a **baseline prior** an analysis is required in which the prior distribution has minimal impact, so that the information in the likelihood dominates the posterior.

The second situation, which we label as a **substantive prior** is one in which it is desired to incorporate more substantial prior information into the analysis.

Baseline Priors

On first consideration it would seem that the specification of a baseline prior is straightforward, one simply takes the choice

$$\pi(\boldsymbol{\theta}) \propto 1 \tag{7}$$

so that the posterior distribution depends solely on the data through the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

There are two difficulties with this:

- May lead to improper posteriors (aka, the end of the world).
- We can't be improper on multiple nonlinear scales.

Improper Priors

The prior $\pi(\boldsymbol{\theta}) \propto 1$ is improper (it does not integrate to a positive constant $< \infty$) unless the range of each element of $\boldsymbol{\theta}$ is finite.

In some instances this is not a problem since the posterior corresponding to the prior is proper.

Philosophically a posterior arising from an improper prior may be justified as a limiting case of proper priors.

More practically we may instead assume that the prior is integrable over its support but is “locally uniform”, so that the likelihood dominates.

Improper Priors

For nonlinear models in particular, care must be taken to ensure that the posterior corresponding to a particular prior choice is proper.

Some general guidelines are available, for example, improper priors for the regression parameters in a generalized linear model will usually lead to a proper posterior although not for some pathological cases.

For example suppose $Y \mid p \sim \text{Binomial}(n, p)$, and a uniform prior is used on the logit of p, $\log\{p/(1-p)\}$ which implies the prior on p is

$$\pi(p) = [p(1-p)]^{-1}.$$

Then an improper posterior results if $y = 0$ (or $y = n$) since the non-integrable spike at $p = 0$ (or $p = 1$) remains in the posterior.

For $n = 1$ one of these events will always occur and so an improper posterior always results.

Improper Priors

To illustrate the non-propriety in another non-linear situation consider the model

$$Y_i \mid \theta \sim_{ind} N\{\exp(-\theta x_i), \sigma^2\}, \tag{8}$$

$i = 1, \ldots, n$, with $\theta > 0$ and σ^2 assumed known.

With an improper uniform prior on θ we have the posterior

$$p(\theta \mid \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - e^{-\theta x_i})^2\right\}.$$

As $\theta \rightarrow \infty$,

$$p(\theta \mid \boldsymbol{y}) \rightarrow \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\},$$

a constant, so that the posterior is improper.

Multiple Scales

The second is that if we reparameterize the model in terms of $\boldsymbol{\phi} = \boldsymbol{g}(\boldsymbol{\theta})$ where $\boldsymbol{g}(\cdot)$ is a one-one mapping, then the prior for $\boldsymbol{\phi}$ corresponding to (7) is given by

$$\pi(\boldsymbol{\phi}) = \left| \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\phi}} \right|,$$

which, unless $\boldsymbol{g}$ is linear, is not constant.

As an example, consider a variance σ^2, the prior $\pi(\sigma^2) \propto 1$ corresponds to a prior for the standard deviation of $\pi(\sigma) \propto \sigma$; the problem is that we cannot be “flat” on different scales.

Multiple Scales

This indicates that a desirable property in constructing baseline priors is there invariance to parameterization, so that we obtain the same prior regardless of the starting parameterization.

In the example just considered suppose the data are normally distributed with variance σ^2.

The improper prior

$$\pi(\sigma) \propto \frac{1}{\sigma}$$

has a number of justifications.

Example: Normal linear regression, variance unknown

Suppose we have $Y_i \mid \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{x}_i\beta, \sigma^2)$, $i = 1, \ldots, n$, $\dim(\boldsymbol{\beta}) = p$.

MLE:

$$\widehat{\boldsymbol{\beta}} \sim t_p(\boldsymbol{\beta}, (\boldsymbol{x}^\top \boldsymbol{x})^{-1} s^2), n - p),$$

a Student t distribution with $n - p$ degrees of freedom.

Improper prior: $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$.

Marginal posterior:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \int p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) d\sigma^2,$$

where

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto l(\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}, \sigma^2).$$

Example: Normal linear regression, variance unknown

Hence,

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}) &= \int \frac{(2\pi\sigma^2)^{-n/2}}{\sigma^2} \exp\left\{-\frac{[(n-p)s^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{x}^\top \boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]}{2\sigma^2}\right\} d\sigma^2 \\ &\propto \int \underbrace{(\sigma^2)^{-(n/2+1)} \exp\left\{-\frac{c}{2\sigma^2}\right\}}_{\text{kernel of an inverse Gamma distribution IGa}(n/2,\, c/2).} d\sigma^2 \end{aligned}$$

where $c = (n-p)s^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{x}^\top \boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Example: Normal linear regression, variance unknown

Therefore,

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}) &\propto \left(\frac{c}{2}\right)^{-n/2} \propto \{(n-p)s^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{x}^\top \boldsymbol{x} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^{-n/2} \\ &\propto \left\{1 + \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{x}^\top \boldsymbol{x} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(n-p)s^2}\right\}^{[-(n-p)+p]/2} \\ &= \left\{1 + \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{n-p}\right\}^{[-(n-p)+p]/2} \end{aligned}$$

where $\Sigma = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} s^2$.

The posterior is,

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim t_p(\widehat{\boldsymbol{\beta}}, (\boldsymbol{x}^\top \boldsymbol{x})^{-1} s^2, n-p).$$

Substantive Priors

We must have a clear understanding of the meaning of the parameters of the model for which we are specifying priors, and this can often be achieved by reparameterization.

It may be easier to specify priors on observable quantities, and then transform back to the parameters.

As a simple example, for the model (8) we might specify a (say, beta) prior for the expected response at $x = \widehat{x}$, $\phi = \exp(-\theta\widehat{x})$ to give a prior $\pi_\phi(\phi)$.

The prior for θ is

$$\pi_\theta(\theta) = \pi_\phi(e^{-\theta\widehat{x}}) \times \widehat{x}e^{-\theta\widehat{x}},$$

the last term corresponding to the Jacobian of the transformation from ϕ to θ.

Substantive Priors

An obvious procedure is to base the prior distribution upon previously collected data.

Preliminary modeling of such data should be carried out to remove the sampling error.

For example, in the simplest case suppose we require a prior for μ where we have data Y_i with $\text{E}[Y_i] = \mu$, $i = 1, \dots, n$, and one has previous data $\boldsymbol{Z} = (Z_1, \dots, Z_m)$.

If one believed that the data-generation mechanism for both sets of data was identical then it would be logical to base the posterior on the combined data.

Often such an assumption cannot be made and a conservative approach is to take the prior as the posterior based on $\boldsymbol{Z}$, with an inflated variance, to accommodate the additional uncertainty.

Example: Lung Cancer and Radon

Recall, the likelihood is

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\left[\ E_i \exp(\beta_0 + \beta_1 x_i)\ \right],$$

where recall that Y_i are counts of lung cancer incidence in Minnesota in 1998–2002, and x_i is a measure of residential radon in county i, $i = 1, \dots, n$.

The obvious improper prior here is $\pi(\boldsymbol{\beta}) \propto 1$ (and results in a proper posterior for this likelihood).

Example: Lung Cancer and Radon

To specify a substantive prior we need to have a clear interpretation of the parameters, and β_0 and β_1 are not the most straightforward to contemplate.

Hence, we reparameterize the model as

$$Y_i \mid \boldsymbol{\theta} \sim_{ind} \text{Poisson}\left(E_i \theta_0 \theta_1^{x_i - \overline{x}}\right),$$

where $\boldsymbol{\theta} = [\theta_0, \theta_1]^{\intercal}$ so that

$$\theta_0 = \mathsf{E}[Y/E \mid x = \overline{x}] = \exp(\beta_0 + \beta_1 \overline{x})$$

is the expected standardized mortality ratio in an area with average radon.

Example: Lung Cancer and Radon

The standardization that leads to expected numbers *E* implies we would expect θ_0 to be centered around 1.

The parameter $\theta_1 = \exp(\beta_1)$ is the relative risk associated with a one-unit increase in radon.

Due to ecological bias, studies often show a negative association between lung cancer incidence and radon.

Example: Lung Cancer and Radon

One convenient choice for positive parameters is the lognormal distribution.

For a generic parameter, θ, denote the prior by $\theta \sim \text{Lognormal}(\mu, \sigma)$.

To obtain the moments of the distribution we may specify the prior median, θ_m, as a "typical" value we would expect, and the 95% point of the prior, θ_u. We then solve for the moments via:

$$\mu = \log(\theta_m), \quad \sigma = \{\log(\theta_u) - \mu\}/1.645.$$

For θ_0 we assume a lognormal prior with 2.5% and 97.5% quantiles of 0.67 and 1.5 to give $\mu = 0, \sigma = 0.21$.

For θ_1 we assume the relative risk associated with a one-unit increase in radon is between 0.8 and 1.2 with probability 0.95, to give $\mu = -0.02, \sigma = 0.10$.

Assuming independence between different parameters often feels shameful, especially when one thinks about reparameterization.

Asymptotic Behavior

Asymptotic Behavior

A "pure" Bayesian approach to inference would dictate that one specifies likelihood and prior and that is sufficient.

But it seems prudent to examine the behavior of the posterior as one hypothetically gathers more data.

Further, the frequentist behavior of Bayes estimators – this assessment is not "required" but seems eminently sensible.

We provide a heuristic development of the asymptotic distribution of the posterior distribution.

We write

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \exp\{\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})\},$$

and let

- θ_0 denote the prior mode, and
- $\widehat{\theta}_n$ the MLE.

Asymptotic Behavior

Taking second-order Taylor series expansions of each of the terms about the modes gives

$$
\begin{aligned}
\log \pi(\boldsymbol{\theta}) &\approx \log \pi(\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\intercal}\boldsymbol{I}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\
\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) &\approx \log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_n) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^{\intercal}\boldsymbol{I}^{\star}(\widehat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)
\end{aligned}
$$

where

$$
\boldsymbol{I}_0 = - \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\intercal}} \log \pi(\boldsymbol{\theta}) \bigg|_{\theta_0} \qquad (9)
$$

$$
\boldsymbol{I}_n^{\star}(\widehat{\boldsymbol{\theta}}_n) = - \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\intercal}} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) \bigg|_{\widehat{\theta}_n} \qquad (10)
$$

is the observed information.

As $n \to \infty$, the likelihood approaches normality, while the prior remains fixed but, as we see, the likelihood dominates.

Asymptotic Behavior

Hence, gathering together the terms in the quadratic forms, the posterior can be written approximately as,

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_n)^\intercal \boldsymbol{J}_n(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_n)\right\},$$

where

$$\begin{aligned} \boldsymbol{J}_n &= \boldsymbol{I}_0 + \boldsymbol{I}_n^\star(\widehat{\boldsymbol{\theta}}_n) \\ \widetilde{\boldsymbol{\theta}}_n &= \boldsymbol{J}_n^{-1}\left\{\boldsymbol{I}_0\boldsymbol{\theta}_0 + \boldsymbol{I}_n^\star(\widehat{\boldsymbol{\theta}}_n)\widehat{\boldsymbol{\theta}}_n\right\} \end{aligned}$$

As $n \to \infty$,

- the observed information tends to the expected information, that is $\boldsymbol{I}^\star(\widehat{\boldsymbol{\theta}}_n) \to \boldsymbol{I}(\widehat{\boldsymbol{\theta}}_n)$, and
- the influence of the prior diminishes, so that

$$\boldsymbol{\theta} \mid \boldsymbol{y} \to_d \mathrm{N}\left(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{I}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}\right).$$

Asymptotic Behavior

We know that the MLE is a consistent estimator if the likelihood is correctly specified.

Since the above shows that $\mathrm{E}_n[\theta \mid \boldsymbol{y}] \rightarrow \widehat{\theta}_n$ as $n \rightarrow n$, we can say that the posterior mean is consistent.

Again, because it mimics the behavior of the MLE the posterior mean is also asymptotically efficient.

The same properties hold for the posterior median.

An important caveat here is that the prior should not exclude any of the parameter space.

Asymptotic Behavior

We might also speculate on the behavior of posterior summaries under model misspecification.

Again we can turn to the behavior of the MLE – if the likelihood is of linear exponential family form, he posterior mean/median are consistent estimators of the parameters in the mean model because the MLEs are.

The spread of the posterior distribution could be completely inappropriate, however.

We re-emphasize that we should always consider the model we feel is most appropriate for the data at hand, but the linear exponential family is advantageous in terms of consistency.

Frequentist Properties of Bayesian Estimators

In terms of frequentist behavior of Bayesian estimators, we briefly describe a simple example to illustrate the trade-offs of prior specification.

Suppose we have Y_i, $i = 1, \dots, n$, with Y_i independently and identically distributed with $\mathrm{E}[\,Y_i \mid \mu] = \mu$ and $\mathrm{var}(\,Y_i \mid \mu) = \sigma^2$ with σ^2 known.

The asymptotic distribution of the MLE is

$$\overline{Y}_n \;\to_d\; \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

We treat this distribution as the likelihood and examine a Bayesian analysis with prior

$$\mu \sim \mathrm{N}(m, v).$$

Frequentist Properties of Bayesian Estimators

The posterior is

$$\mu \mid \overline{Y} \;\to\; {}_d\mathrm{N}\left(w\overline{Y} + (1-w)m, w\frac{\sigma^2}{n} \right)$$

where

$$w = \frac{nv}{nv + \sigma^2}.$$

We first observe that the posterior mean is consistent so long as $v > 0$.

The mean squared error (MSE) of the posterior mean is,

$$\begin{aligned} \text{MSE} &= \text{Variance} + \text{Bias}^2 \\ &= w\frac{\sigma^2}{n} + \{w\mu + (1-w)m - \mu\}^2. \end{aligned}$$

Frequentist Properties of Bayesian Estimators

Figure 4 illustrates the MSE as a function of μ for two different prior disrtributions that are both centered at zero.

The trade-off when specifying the variance of the prior is clear; if μ is close to zero then greater gains in MSE are achieved with a small v though the range of μ over which an improved MSE is achieved is narrower.

At values of μ of $m \pm \sqrt{v + \sigma^2/n}$ the MSE of the MLE and Bayes estimator are equal.

The variance of the estimator is given by the lowest point of the MSE curves so that we see that the bias dominates for large $\mid \mu \mid$.

MSE of Bayes Estimator

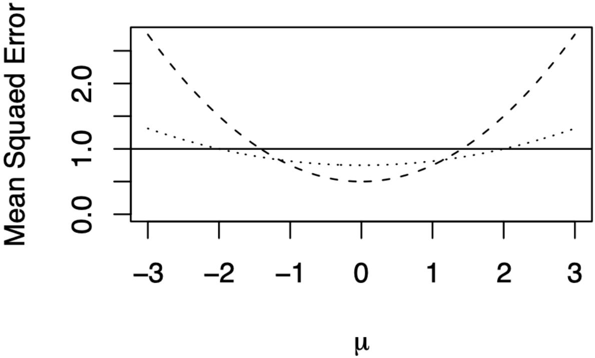


Figure 4: Mean square error of posterior mean with $v = 1$ (dashed line) and $v = 3$ (dotted line) as a function of the parameter μ. The mean squared error of the MLE is the solid horizontal line.

Computation

Implementation

Historically, the inability of users to carry out the required integrations, was a serious barrier to widespread use of Bayesian methods.

Frequentist methods are typically way easier to implement requiring:

- Maximization/root finding for point estimates and second derivatives for asymptotic inference.
- Resampling methods, which are often straightforward and automatic to implement without any great ingenuity.

Bayes computation has taken great strides in the last 35 years, starting with the MCMC revolution, and then the re-emergence of Laplace-based methods through INLA and related approaches.

Conjugate Analysis

So-called conjugate prior distributions allow analytical evaluation of many of the integrals required for Bayesian inference, at least for certain convenient parameters.

A conjugate prior is such that $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(\boldsymbol{\theta})$ belong to the same family, though this definition is not adequate since it will always be true given a suitable definition of the family of distributions.

The following is notation-heavy, but often in practice one can quickly recognize conjugate priors and turn the Bayesian handle.

To obtain a more useful class we first note that if $\boldsymbol{T}(\boldsymbol{Y})$ denotes a **sufficient statistic** for a particular likelihood $p(\cdot|\boldsymbol{\theta})$, then

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{\theta}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

This allows a definition of a conjugate family in terms of likelihoods that admit a sufficient statistic of fixed dimension.

The exponential family of distributions

Exponential family distributions $\mathcal{F}$ have the form:

$$p(y_i|\boldsymbol{\theta}) = f(y_i)g(\theta)\exp\{\phi(\boldsymbol{\theta})^\intercal \boldsymbol{u}(y_i)\},$$

where, in general, $\phi(\boldsymbol{\theta})$ and $\boldsymbol{u}(y_i)$ have the same dimension as $\boldsymbol{\theta}$, and $\phi(\boldsymbol{\theta})$ is called the natural parameter of $\mathcal{F}$.

For *n* iid observations from $p(\cdot|\boldsymbol{\theta})$:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \left[\prod_{i=1}^{n} f(y_i)\right] g(\boldsymbol{\theta})^n \exp\{\phi(\boldsymbol{\theta})^\intercal \boldsymbol{t}(\boldsymbol{y})\},$$

where

$$\boldsymbol{t}(\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{u}(y_i).$$

The conjugate prior density is then defined as:

$$dredp(\boldsymbol{\theta}) = c(\eta, \boldsymbol{\upsilon}) \times g(\boldsymbol{\theta})^\eta \exp\{\phi(\boldsymbol{\theta})^\intercal \boldsymbol{\upsilon}\},$$

where η and $\boldsymbol{\upsilon}$ are specified, a priori.

Posterior and Predictive Distributions

The posterior distribution is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = c(\eta + n, \boldsymbol{\upsilon} + \boldsymbol{t}) \times g(\boldsymbol{\theta})^{\eta+n} \exp\{\phi(\boldsymbol{\theta})^{\intercal}[\boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y})]\},$$

demonstrating **conjugacy.**

Comparison with $p(y_i|\boldsymbol{\theta})$ indicates that:

- η may be viewed as a prior sample size,
- giving rise to a sufficient statistic $\boldsymbol{\upsilon}$.

Posterior and Predictive Distributions

Suppose now that we are interested in obtaining the predictive distribution for new observations

$$\boldsymbol{Z} = (Z_1, \ldots, Z_m)$$

arising as an iid sample from $p(\cdot|\boldsymbol{\theta})$.

In the case of a conjugate prior:

$$\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{y}) &= \int_\theta p(\boldsymbol{z}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \\
&= \int_\theta \left[\prod_{i=1}^{m} f(z_i)\right] g(\boldsymbol{\theta})^m \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\intercal \boldsymbol{t}(\boldsymbol{z})\} \times c(\eta + n, \boldsymbol{\upsilon} + \boldsymbol{t}) \\
&\times \; g(\boldsymbol{\theta})^{\eta+n} \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\intercal [\boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y})]\}, \\
&= \left\{\prod_{i=1}^{m} f(z_i)\right\} \frac{c(\eta + n, \boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y}))}{c(\eta + n + m, \boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y}, \boldsymbol{z}))}.
\end{aligned}$$

Example: Binomial likelihood

In exponential family form,

$$p(y|\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) (1-\theta)^n \exp\left\{ y \log \frac{\theta}{1-\theta} \right\}.$$

The conjugate prior is therefore identified as:

$$\begin{aligned} p(\theta) &= c(\eta, \upsilon)(1-\theta)^\eta \exp\left\{ \upsilon \log \frac{\theta}{1-\theta} \right\} \\ &= \frac{\Gamma(\eta+2)}{\Gamma(\upsilon+1)\Gamma(\eta-\upsilon+1)} \theta^\upsilon (1-\theta)^{\eta-\upsilon}, \end{aligned}$$

the beta distribution $\mathrm{Be}(a = \upsilon + 1, b = \eta - \upsilon + 1)$.

Sample size $\eta = a + b - 2$ yields the prior sufficient statistic $\upsilon = a - 1$.

It follows immediately that the posterior is

$$\mathrm{Be}(a + y, b + n - y).$$

Example: Binomial likelihood

We have

$$\begin{aligned} \mathrm{E}[\theta|y] &= \frac{a+y}{a+b+n} \\ &= \frac{y}{n}\frac{n}{a+b+n} + \frac{a}{a+b}\frac{a+b}{a+b+n}. \end{aligned}$$

Note that $a = b = 1$ does not give the MLE – why???

We would rarely report the mode, but it offers some insight:

$$\begin{aligned} \mathrm{mode}[\theta|y] &= \frac{a+y-1}{a+b+n-2} \\ &= \frac{y}{n}\frac{n}{a+b+n-2} + \frac{a-1}{a+b-2}\frac{a+b-2}{a+b+n-2}. \end{aligned}$$

The prior choice $a = b = 1$ results in the posterior mode equalling the MLE, as expected.

Example: Binomial likelihood

The marginal distribution of the data, given likelihood and prior, is

$$p(y) = \Pr(Y = y) = \left(\begin{array}{c} n \\ y \end{array} \right) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \times \frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a+b+n)},$$

$y = 0, \dots, n$.

If $a = b = 1$,

$$p(y) = (n+1)^{-1}$$

for $y = 0, 1, \dots, n$, is uniform.

Obvious???

Example: Binomial likelihood

In general, the mean and variance of the prior predictive are

$$\begin{aligned} \mathsf{E}[Y] &= \mathsf{E}_\theta\{\mathsf{E}[Y|\theta]\} = n \times \mathsf{E}[\theta] = n \times \frac{a}{a+b}, \\ \text{var}(Y) &= n \times \mathsf{E}[\theta](1 - \mathsf{E}[\theta]) \times \frac{a+b+n}{a+b+1}, \end{aligned}$$

illustrating the overdispersion relative to $p(Y|\theta_\mathsf{F})$ with θ_F some fixed number.

Overdispersion is not possible if $n = 1$ (a Bernoulii is a Bernoulli and there are no other possibilities for binary data).

This is of interest since we can specify priors and examine their implications in terms of the observables we expect to see – often it will be more straightforward to think in terms of the observables rather than model parameters.

In general we may simulate data from our prior (if proper!) and examine these data to see if they correspond to our prior beliefs.

Notes on Conjugacy 1

- Whether a baseline prior or a subjective prior is to be chosen, the conjugate family may be utilized. In some cases the limiting form of the conjugate family corresponds to a frequentist analysis (e.g., normal likelihood, unknown mean, known variance).

- We have seen that for certain parameters inference is straightforward. For functions of interest, however, the required integrals are not available in closed form. In this case a sampling-based approach (e.g., direct sampling, see later) may be utilized – conjugate families are straightforward to sample from.

- Although the number of models for which conjugate families exist is limited (particularly for regression), we shall see that when Markov chain Monte Carlo approaches are considered, the conditional forms required may be of conjugate form.

Notes on Conjugacy 2

- In general caution should be exercised in selecting a prior distribution merely for computational convenience.

- Multivariate conjugate priors are often deficient in terms of the flexibility they might provide since they do not have sufficient parameters to specify, e.g., the normal, Wishart, dirichlet distributions.

- When we specify the values of the constants occurring in the prior we may equate prior guesses of moment summaries of the parameter to the corresponding moments of the prior.

  Example: if $\theta \sim \text{Ga}(a, b)$ then we have $\text{E}[\theta] = a/b$ and $\text{var}(\theta) = a/b^2$. If our prior guesses at the first two moments are m and v then we may assign $a = m^2/v$ and $b = m/v$.

- Quantiles may be equated also and this is usually more natural than moments.

Choosing a Prior: Approach One

- Recall that to specify a beta distribution, we need to specify two quantities, *a* and *b*, which are difficult to interpret.
- The posterior mean is

$$\mathrm{E}[\theta|y] = \frac{y+a}{N+a+b} = \frac{y}{N}\underbrace{\frac{N}{N+a+b}}_{W} + \frac{a}{a+b}\underbrace{\frac{a+b}{N+a+b}}_{1-W}.$$

- Viewing the denominator as a sample size suggests a method for choosing *a* and *b*.
- We may specify the prior mean $m_{\text{prior}} = a/(a+b)$ and the "prior sample size" $N_{\text{prior}} = a+b$
- We then solve for *a* and *b* via

$$\begin{aligned} a &= N_{\text{prior}} \times m_{\text{prior}} \\ b &= N_{\text{prior}} \times (1 - m_{\text{prior}}). \end{aligned}$$

- Intuition: *a* is like a prior number of successes and *b* like the prior number of failures.

A Binomial Example

- Suppose we set $N_{\text{prior}} = 5$ and $m_{\text{prior}} = \frac{2}{5}$.
- It is as if we saw 2 successes out of 5.
- Suppose we obtain data with $y = 7$, $N = 10$ and so $\frac{y}{N} = \frac{7}{10}$.
- Hence $\mathrm{W} = 10/(10+5)$ and

$$\begin{aligned} \mathrm{E}[\theta|y] &= \frac{7}{10} \times \frac{10}{10+5} + \frac{2}{5} \times \frac{5}{10+5} \\ &= \frac{9}{15} = \frac{3}{5}. \end{aligned}$$

- Solving:

$$\begin{aligned} a &= N_{\text{prior}} \times m_{\text{prior}} = \qquad 5 \times \frac{2}{5} = 2 \\ b &= N_{\text{prior}} \times (1 - m_{\text{prior}}) = 5 \times \frac{3}{5} = 3 \end{aligned}$$

- This gives a $\text{Beta}(y + a, N - y + b) = \text{Beta}(7 + 2, 3 + 3)$ posterior.

Beta Prior, Likelihood and Posterior

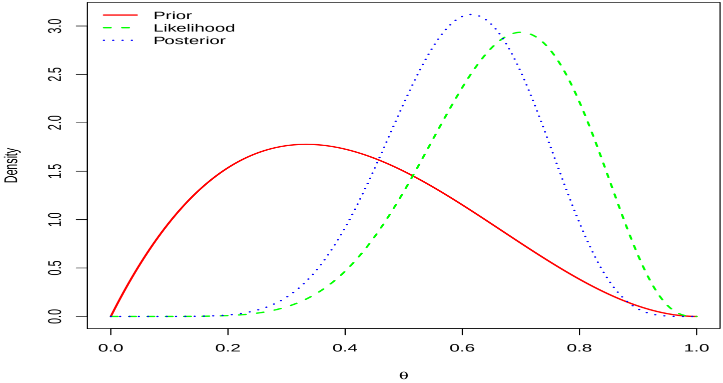


Figure 5: The prior is Beta(2,3) the likelihood is proportional to a Beta(7,3) and the posterior is Beta(7+2,3+3).

Choosing a Prior: Approach Two

- An alternative convenient way of choosing a and b is by specifying two quantiles for θ with associated (prior) probabilities.
- For example, we may wish $\Pr(\theta < 0.1) = 0.05$ and $\Pr(\theta > 0.6) = 0.05$.
- The values of a and b may be found numerically. For example, we may solve

$$[p_1 - \Pr(\theta < q_1 | a, b)]^2 + [p_2 - \Pr(\theta < q_2 | a, b)]^2 = 0$$

for a, b.

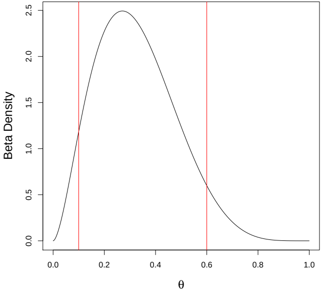


Figure 6: Beta(2.73,5.67) prior with 5% and 95% quantiles highlighted.

Laplace Approximations

Laplace approximations have a long history in statistics, and are still finding use!

Let,

$$I = \int \exp\left[\, ng(\theta)\, \right] d\theta,$$

denote a generic integral of interest and suppose:

- $\widetilde{\theta}$ is the maximum of $g(\cdot)$.

Consider the Taylor series expansion,

$$ng(\theta) = n \sum_{k=0}^{\infty} \frac{(\theta - \widetilde{\theta})^k}{k!} g^{(k)}(\widetilde{\theta}),$$

where $g^{(k)}(\widetilde{\theta})$ represents the k-th derivative of $g(\cdot)$ evaluated at $\widetilde{\theta}$.

Laplace Approximations

Hence, we can write the required integral as,

$$\begin{aligned} I &= \int \exp\left[\; n \sum_{k=0}^{\infty} \frac{(\theta-\tilde{\theta})^k}{k!} g^{(k)}(\tilde{\theta}) \;\right] d\theta \\ &= e^{ng(\tilde{\theta})} \int \exp\left[\; \frac{(\theta-\tilde{\theta})^2}{2/[ng^{(2)}(\tilde{\theta})]} \;\right] \exp\left[\; n \sum_{k=3}^{\infty} \frac{(\theta-\tilde{\theta})^k}{k!} g^{(k)}(\tilde{\theta}) \;\right] d\theta \end{aligned}$$

Taking the approximation to the second term of the Taylor series gives the estimate,

$$\widehat{I} = \exp\left[\; ng(\tilde{\theta}) \;\right] \left(\frac{2\pi v}{n}\right)^{1/2},$$

where

$$v = -1/[g^{(2)}(\tilde{\theta})],$$

is incorporating the curvature at the maximum.

In asymptotic terms, Laplace's method typically has an error of order $O(n^{-1})$.

Laplace in a Bayesian context

In a Bayesian context, suppose we wish to evaluate the posterior expectation of a positive function of interest $\phi(\theta)$, i.e.,

$$\begin{aligned}\mathsf{E}[\phi(\theta) \mid \boldsymbol{y}] &= \frac{\int \exp\left[\, \log\phi(\theta) + \log p(\boldsymbol{y} \mid \theta) + \log\pi(\theta) \,\right] \, d\theta}{\int \exp\left[\, \log p(\boldsymbol{y} \mid \theta) + \log\pi(\theta) \,\right] \, d\theta} \\ &= \frac{\int \exp\left[\, n g_1(\theta) \,\right] \, d\theta}{\int \exp\left[\, n g_2(\theta) \,\right] \, d\theta}.\end{aligned}$$

Application of Laplace's method to numerator and denominator gives

$$\widehat{\mathsf{E}}[\phi(\theta) \mid \boldsymbol{y}] = \frac{\widetilde{v}_1}{\widetilde{v}_2} \frac{\exp\left[\, n g_1(\widetilde{\theta}_1) \,\right]}{\exp\left[\, n g_2(\widetilde{\theta}_2) \,\right]}$$

where $\widetilde{\theta}_j$ is the maximum of $g_j(\cdot)$ and $\widetilde{v}_j = -1/g_j^{(2)}(\widetilde{\theta}_j)$, $j = 1, 2$.

Laplace in a Bayesian context

It may be shown that

$$\widehat{\mathsf{E}}[\phi(\theta) \mid \boldsymbol{y}] = \mathsf{E}[\phi(\theta) \mid \boldsymbol{y}](1 + O(n^{-2})),$$

since errors in the numerator and denominator cancel.

If ϕ is not positive then a simple solution is to add a large constant to ϕ; Laplace's method may then be applied with the constant subtracted at the end.

See Tierney and Kadane (1986) for more details in a Bayesian context.

Direct Sampling

Given independent samples $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, m\}$ with

$$\boldsymbol{\theta}^{(t)} = [\theta_1^{(t)}, \ldots, \theta_p^{(t)}]$$

from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ the univariate marginal posterior for $p(\theta_j \mid \boldsymbol{y})$ may be represented by histograms or density estimated constructed from the points $\theta_j^{(t)}$, $t = 1, \ldots, m$.

Similarly bivariate marginals can be visualized using scatterplots.

Posterior means $\mathrm{E}[\theta_j \mid \boldsymbol{y}]$ may be approximated by

$$\widehat{\mathrm{E}}[\theta_j \mid \boldsymbol{y}] = \frac{1}{m} \sum_{t=1}^{m} \theta_j^{(t)},$$

with other moments following in an obvious fashion.

Coverage probabilities of the form $\Pr(a < \theta_j < b \mid \boldsymbol{y})$ are estimated by

$$\widehat{\Pr}(a < \theta_j < b \mid \boldsymbol{y}) = \frac{1}{m} \sum_{t=1}^{m} 1_{[a < \theta_j^{(t)} < b]}.$$

Difference in Binomial Proportions

- Savage *et al.* (2008) give data on allele frequencies within a gene that has been linked with skin cancer.
- It is interest to examine differences in allele frequencies between populations.
- We examine one SNP and extract data on Northern European (NE) and United States (US) populations.
- Let θ_1 and θ_2 be the allele frequencies in the NE and US population from which the samples were drawn, respectively.
- The allele frequencies were 10.69% and 13.21% with sample sizes of 650 and 265, in the NE and US samples, respectively.
- We assume independent Beta(1,1) priors on each of θ_1 and θ_2.
- The posterior probability that $\theta_1 - \theta_2$ is greater than 0 is 0.12 (computed as the proportion of the samples $\theta_1^{(s)} - \theta_2^{(s)}$ that are greater than 0), so there is little evidence of a difference in allele frequencies between the NE and US samples.

Binomial Two Sample Example

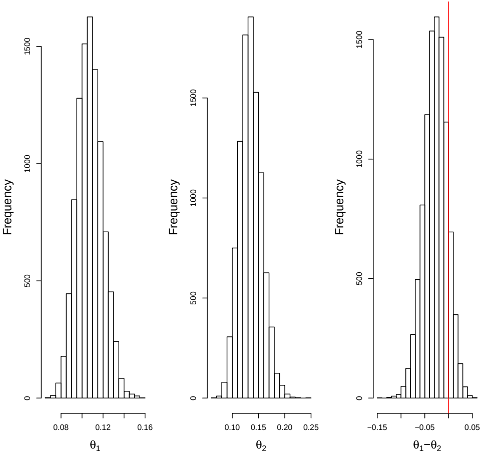


Figure 7: Histogram representations of $p(\theta_1|y_1)$, $p(\theta_2|y_2)$ and $p(\theta_1 - \theta_2|y_1, y_2)$. The red line in the right plot is at the reference point of zero.

Direct Sampling

We briefly describe a rejection algorithm that can be used to generate samples from the posterior.

Let $\boldsymbol{\theta}$ denote the unknown parameters and then obtain the maximized likelihood (which will often be stratightforward)

$$M = \sup_{\theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}})$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE. The algorithm then proceeds as follows:

1. Generate $U \sim U(0,1)$ and, independently, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.
2. Accept $\boldsymbol{\theta}$ if
$$U < \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{M} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}})},$$
otherwise return to 1.

The probability that a point is accepted is,

$$p_a = \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{M} = \frac{p(\boldsymbol{y})}{M}.$$

An Example

Suppose a seroprevalence test is carried out with sensitivity

$$\delta = \text{Pr(+ve test} \mid \text{disease)}$$

and specificity,

$$\gamma = \text{Pr(-ve test} \mid \text{no disease).}$$

Let π be the true prevalence.

We test *n* people and *y* are recorded as having the disease, and a starting model is

$$y|p \sim \text{Binomial}(N, p)$$

where *p* is the probability of a +ve test result. with

$$\begin{aligned} p &= \text{Pr(+ve test)} \\ &= \text{Pr(+ve test} \mid \text{disease) Pr(disease)} \\ &+ \text{Pr(+ve test} \mid \text{no disease) Pr(no disease)} \\ &= \delta\pi + (1-\gamma)(1-\pi) \\ &= \pi(\delta+\gamma-1) + (1-\gamma) \end{aligned}$$

An Example

Suppose for simplicity the sensitivity and specificity are known and we want to estimate π.

With this binomial model the MLE is (exercise!):

$$\widehat{\pi} = \frac{y - N(1-\gamma)}{N(\delta + \gamma - 1)}$$

A Bayesian model is

$$\begin{aligned} y|\pi &\sim \text{Binomial}(N, \pi(\delta + \gamma - 1) + (1-\gamma)) \\ \pi &\sim \text{Beta}(a, b) \end{aligned}$$

Not conjugate!

However, the simple rejection algorithm can be implemented that simulates samples from the posterior $p(\pi|y)$.

COVID-19 Prevalence Estimate

- In early April, 2020, Bendavid *et al.* (2020) recruited 3330 residents of Santa Clara County, California and tested them for COVID-19 antibodies. 50 people tested positive, yielding a raw estimate of 1.50%.
- We take the sensitivity as 0.8 and the specificity as 0.995 and the prior parameters as $a = b = 1$.
- See Gelman and Carpenter (2020) for a more comprehensive Bayesian analysis.

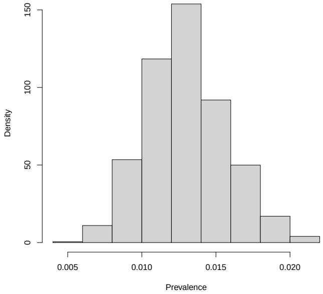


FIGURE 8: Histogram representation of the posterior distribution for the prevalence π. The posterior median is 1.28% and a 90% interval is (0.88%,1.77%).

Gaussian Quadrature

A general method of integration is provided by quadrature (numerical integration) in which an integral

$$I = \int f(u)\,du,$$

is approximated by

$$\widehat{I} = \sum_{i=1}^{n_w} f(u_i) w_i,$$

for

- design points $u_1, \ldots, u_{n_w}$ and
- weights $w_1, \ldots, w_{n_w}$.

Different choices of (u_i, w_i) lead to different integration rules.

Gaussian Quadrature

Fundamental Theorem of Gaussian Quadrature:

The abscissas of the *N*-point Gaussian quadrature formula are precisely the roots of the orthogonal polynomial for the same interval and weighting function.

In Bayesian applications we have integrals which, in large samples in particular, are with respect to a normal density.

Gauss-Hermite quadrature is designed for problems of this type.

Gaussian Quadrature

Specifically, it provides exact integration of

$$\int_{-\infty}^{\infty} g(u)e^{-u^2}\, du,$$

where $g(\cdot)$ is a polynomial of degree $2n_w - 1$.

The design points are the zeroes of the so-called Hermite polynomials.

Specifically, for a rule of n_w points, u_i is the $i-$th zero of $H_{n_w}(u)$, the Hermite polynomial of degree n_w, and

$$w_i = \frac{w^{n_w - 1} n_w! \sqrt{\pi}}{n_w^2 [H_{n_w - 1}(u_i)]^2}.$$

To implement this method the function must be centered and scaled in some way, for example we could center and scale by the current estimates of the mean and standard deviation – known as adaptive quadrature.

Gaussian Quadrature

Now suppose $\boldsymbol{\theta}$ is two-dimensional and we wish to evaluate

$$\begin{aligned} I &= \int f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int\int f(\theta_1, \theta_2)\, d\theta_2 d\theta_1 \\ &= \int f^\star(\theta_1)\, d\theta_1, \end{aligned}$$

where

$$f^\star(\theta_1) = \int f(\theta_1, \theta_2)\, d\theta_2.$$

Gaussian Quadrature

We form

$$\widehat{I} = \sum_{i=1}^{m_1} w_i \widehat{f}^*(\theta_{1i}),$$

where

$$\widehat{f}^*(\theta_{1i}) = \sum_{j=1}^{m_2} u_j f(\theta_{1i}, \theta_{2j}).$$

Then we have the approximation,

$$\widehat{I} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_i u_j f(\theta_{1i}, \theta_{2j}),$$

which is known as the Cartesian Product.

Scaling and reparameterization are important.

Suppose we know that the posterior mean vector and variance-covariance matrix are given by $\boldsymbol{m}$ and $\boldsymbol{V}$.

We then form,

$$\boldsymbol{x} = \boldsymbol{L}(\theta - \boldsymbol{m})$$

where $\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{V}^{-1}$ and carry out integration in the space of the random variables $\boldsymbol{X}$.

There is no guarantee that the most efficient rule is obtained by scaling in terms of the posterior mean and variance, but we note that the ‘best’ normal approximation to a density (in terms of Kullbach-Leibler divergence) has the same mean and variance.

Importance Sampling

Rather than deterministically selecting points we may randomly generate points from some density $h(\boldsymbol{\theta})$.

We have

$$I = \int_0^1 f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \frac{f(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} h(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathsf{E}_h[w(\boldsymbol{\theta})],$$

where the importance sampling weights:

$$w(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}.$$

A good candidate density $h(\cdot)$ is straightforward to generate samples from, and should "look like" $f(\cdot)$, which we discuss in more detail shortly.

Importance Sampling

Hence, we have the obvious estimator

$$\widehat{I} = \frac{1}{m} \sum_{i=1}^{m} w(\boldsymbol{\theta}_i),$$

where

$$\theta_i \sim_{iid} h(\cdot).$$

We have $\mathsf{E}[\widehat{I}] = I$ and

$$V = \text{var}\left(\widehat{I}(\boldsymbol{\theta})\right) = \frac{1}{m}\text{var}\left(w(\boldsymbol{\theta})\right).$$

From this expression it is clear that a good $h(\cdot)$ will result in $w(\boldsymbol{\theta})$ being approximately constant.

Importance Sampling

The variance can be estimated using the same samples.

In particular, we may estimate V via

$$\widehat{V} = \frac{1}{m}\sum_{i=1}^{m}\frac{f^2(\boldsymbol{\theta}_i)}{h^2(\boldsymbol{\theta}_i)} - \frac{1}{m}\widehat{I}^2,$$

and (appealing to the central limit theorem) $\widehat{I}$ is asymptotically normal and so a $100(1-\alpha)\%$ confidence interval is

$$\widehat{I} \pm Z_{\alpha/2}\widehat{V}^{1/2},$$

where $Z_{\alpha/2}$ is the $\alpha/2$ point of an $\mathrm{N}(0,1)$ random variable.

Hence, the accuracy of the approximation may be directly assessed, providing an advantage over analytical approximations and quadrature methods.

Notes on Importance Sampling

- We require an $h(\cdot)$ with heavier tails than the integrand. We can carry out importance sampling with any h but if the tails are lighter we will have an estimator with infinite variance.
- Many suggestions for $h(\cdot)$ have been made including Student's t distributions and mixtures of Student's t distributions.
- Iteration (for example centering and scaling in a "good" place) may again be used to obtain an estimator with good properties.

Notes on Implementation

- If the number of parameters is small then numerical integration techniques (e.g. quadrature) are highly efficient in terms of the number of function evaluations required. Hence, if, for example, obtaining a point on the likelihood surface is computationally expensive (as occurs if a large simulation is required) then such techniques are preferable to Monte Carlo methods.
- The method employed will depend on whether it is for a one-off application, in which case ease-of-implementation is a consideration, or for a great deal of use, in which case an efficient method may be required.
- In general it is difficult to assess the accuracy of Laplace/numerical integration techniques.
- For simulation methods, independent samples ideal for assessing MC error since s.e.'s on expectations of interest are simply calculated.

Example: Poisson likelihood, normal Prior

We illustrate some of the technique described in this section using a simple example with a Poisson likelihood.

Seascale is a village 3km to the south of Sellafield and had four cases of lymphoid malignancy among 0–14 year olds during 1968–82, compared with 0.25 expected cases (based on the number of children in the region and registration rates for the overall Northern region of England).

A question here is whether such a large number of cases could have reasonably occurred by chance.

Example: Poisson likelihood, normal Prior

We assume the model

$$Y \mid \theta \sim \text{Poisson}\,(\; E \exp(\theta) \;)\,,$$

where e^{θ} is the relative risk.

The MLE is $\widehat{\theta} = \log 16 = 2.77$ with asymptotic standard error 0.25.

We assume an $N(a, b^2)$ normal prior for θ, i.e., a lognormal prior for the relative risk $\exp(\theta)$.

To choose the prior parameters we assume that the median relative risk is 1, and the 90% point of the prior is 10, which leads to $a = 0$ and $b = 1.38$.

Example: Poisson likelihood, normal Prior

We take as aim the estimation of,

$$
\begin{aligned}
I^r &= \int_{-\infty}^{\infty} \theta^r \Pr(y \mid \theta)\pi(\theta)\, d\theta \\
&= \frac{E^y(2\pi b^2)^{-1/2}}{y!} \int \exp\left[r \log\theta - Ee^{\theta} + \theta y - \frac{(\theta - a)^2}{2b^2} \right]\, d\theta \\
&= \frac{E^y(2\pi^2 b^2)^{-1/2}}{y!} \int \exp\left[g_r(\theta) \right]\, d\theta
\end{aligned}
$$

for $r = 0, 1, 2$ to give the normalizing constant, mean and variance via, respectively,

$$
\begin{aligned}
p(\boldsymbol{y}) &= I^0 \\
\mathrm{E}[\theta \mid y] &= \frac{I^1}{I^0} \\
\mathrm{var}(\theta \mid y) &= \frac{I^2}{I^0} - \left(\frac{I^1}{I^0} \right)^2
\end{aligned}
$$

Example: Poisson likelihood, normal Prior

Recall Laplace's method:

$$\begin{aligned}
\widehat{I}^r &= \exp\left[\; n g_r(\widetilde{\theta}) \;\right] \left(\frac{2\pi v}{n} \right)^{1/2}, \\
&= \exp\left[n g_r(\widetilde{\theta}) \right] \left(\frac{2\pi [g_r^{(2)}(\widetilde{\theta})]}{n} \right)^{1/2}.
\end{aligned}$$

In our example,

$$\begin{aligned}
g_r(\theta) &= r \log \theta - E e^{\theta} + \theta y - \frac{(\theta - a)^2}{2b^2} \\
g_r^{(1)}(\theta) &= \frac{r}{\theta} - E e^{\theta} + y - \frac{\theta - a}{b^2} \\
g_r^{(2)}(\theta) &= -\frac{r}{\theta^2} - E e^{\theta} \frac{1}{b^2},
\end{aligned}$$

for $r = 0, 1, 2$.

The results of the Laplace approximation are shown in Table 1, which also shows the results from a Gauss-Hermite rule using 5 points, and centered and scaled by the Laplace approximations to the mean and

Example: Poisson likelihood, normal Prior

The results of the Laplace approximation are shown in Table 1, which also shows the results from a Gauss-Hermite rule using 5 points, and centered and scaled by the Laplace approximations to the mean and standard deviation of the posterior.

	Laplace Approximation	Gauss-Hermite $m = 5$	Importance Sampling	Rejection Algorithm
$\Pr(y)$ $(\times 10^3)$	1.35	1.36	1.36 (1.35,1.37)	1.37
$E[\theta \mid y]$	2.29	2.27	2.27 (2.25,2.29)	2.27 (2.25,2.28)
$\text{var}(\theta \mid y)$	0.304	0.328	0.313 (0.239,0.387)	0.343 (0.176,0.510)

Table 1: Laplace, Gauss-Hermite and Monte Carlo approximations for Poisson-Normal model, both Monte Carlo methods using $m = 500$ points. Truth: 1.37 $(\times 10^3)$, 2.27, 0.329.

Example: Poisson likelihood, normal Prior

We now turn to importance sampling.

As proposal we use a normal distribution with mean and variance given by $\widetilde{\theta}$ and $\widetilde{v}$ from the Laplace approximation.

Table 1 shows estimates resulting from $m = 500$.

The delta method can be used to produce measures of accuracy for functions of $\widehat{I}^r$. For example

$$\text{var}\left(\frac{\widehat{I}^1}{\widehat{I}^0}\right) \approx \frac{\text{var}(\widehat{I}_1)}{\widehat{I}_0^2} + \frac{\widehat{I}_1^2 \text{var}(\widehat{I}_0^2)}{\widehat{I}_4^0}$$

from which asymptotic confidence intervals for the estimate of the posterior mean may be obtained.

Example: Poisson likelihood, normal Prior

Finally we implement a rejection algorithm, sampling from the prior distribution.

The empirical rejection rate can be used to derive the normalizing constant as

$$\widetilde{p}(\boldsymbol{y}) = M \times \widehat{p}_a \tag{11}$$

The variance of this estimator is calculated using the negative binomial variance.

The acceptance probability was 0.07, the small value being explained by the discrepancy between the prior and the likelihood.

Figure 9 gives histogram representations of $p(\theta \mid y)$ using samples from the rejection algorithm.

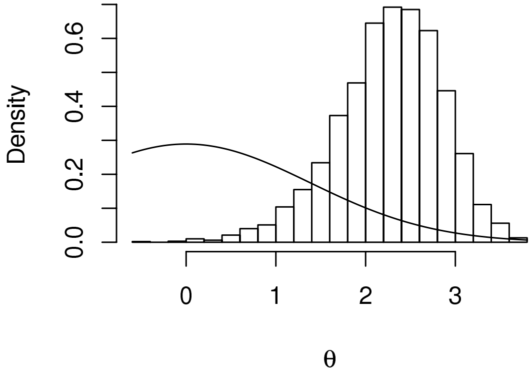


Figure 9: Histogram representation of posterior distribution for θ in the Sellafield example, with prior superimposed as a solid line.

Integrated Nested Laplace Approximation (INLA)

- INLA the method is becoming increasingly popular as a Bayesian computational tool, in large part because of the `INLA` `R` package.

- INLA combines Laplace approximations
  and numerical integration in a very efficient manner – first introduced in Rue *et al.* (2009).

- The method is designed for latent Gaussian models (LGMs).

- Suppose the model has the form

$$\begin{aligned} y_i|\boldsymbol{x}_i, \theta_1 &\sim \text{Likelihood Function} \\ \boldsymbol{x}|\theta_2 &\sim \text{N}(\mathbf{0}, \boldsymbol{Q}(\theta_2)^{-1}) \end{aligned}$$

  where $\boldsymbol{x}$ denotes a vector of variables with normal priors, for example, regression coefficients and random effects and θ_1 and θ_2 are variance components.

Integrated Nested Laplace Approximation (INLA)

- We also have a prior, $\pi(\boldsymbol{\theta})$, for $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$ — these priors are non-normal, because the variance component parameters are usually not on the real line.

- The posterior has the form:

$$
\begin{aligned}
\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \quad &\propto \quad \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}_2) \prod_i p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_1) \\
&\propto \quad \pi(\boldsymbol{\theta}) \mid \boldsymbol{Q}(\boldsymbol{\theta}_2) \mid^{p/2} \exp\left\{ -\frac{1}{2} \boldsymbol{x}^{\mathsf{T}} \boldsymbol{Q}(\boldsymbol{\theta}_2) \boldsymbol{x} + \sum_i \log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_1) \right\}
\end{aligned}
$$

INLA

INLA calculates the univariate posteriors marginals:

$$
\begin{aligned}
\pi(\theta_j|\boldsymbol{y}) &= \int\int \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{x}d\boldsymbol{\theta}_{-j} && (12)\\
&= = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta}_{-j} && (13)\\
\pi(x_i|\boldsymbol{y}) &= \int\int \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{x}_{-i}d\boldsymbol{\theta}\\
&= \int \left[\int \pi(x_i,\boldsymbol{x}_{-i}|\boldsymbol{\theta},\boldsymbol{y})d\boldsymbol{x}_{-i}\right] \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta}\\
&= \int \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta} && (14)
\end{aligned}
$$

INLA

The latent field $\boldsymbol{x}$ and the variance components $\boldsymbol{\theta}$ are treated quite differently by INLA, because the latter are less normal-like in general, even after reparameterization.

The nested part of INLA reflects that given values of $\boldsymbol{\theta}$ Laplace approximations are carried out for $\boldsymbol{x}$, and these are averaged over using numerical integration techniques.

We now describe the various approximations used in INLA.

INLA

The marginal posterior for $\boldsymbol{\theta}$ is, for any value of $\boldsymbol{x}$,

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\end{aligned}$$

The numerator is available, while the denominator is in general not.

The approximation is,

$$\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})\propto\frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^k)p(\boldsymbol{x}|\boldsymbol{\theta}^k)\pi(\boldsymbol{\theta}^k)}{\widehat{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}^k,\boldsymbol{y})} \tag{15}$$

where $\widehat{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}^k,\boldsymbol{y})$ is the Gaussian approximation to the conditional which is obtained by matching the mode and the curvature at the mode.

INLA

The marginal (14), i.e., $\pi(x_i|\boldsymbol{y})$, needs to be calculated for a potentially very long vector $\boldsymbol{x}$.

We could take the marginal from $\widehat{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}^k, \boldsymbol{y})$ but unfortunately this is not generally very accurate.

As an alternative, rewrite as

$$\begin{aligned}\pi(x_i|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}\end{aligned}$$

and the denominator can again be estimated estimated using the Tierney and Kadane (1986) density approximation.

Rue *et al.* (2009) describe a third approximation, the simplified Laplace which corrects the Gaussian approximation for location and skewness using a Taylor's series expansion about the mode.

INLA

The INLA computing scheme therefore consists of (Martino and Riebler, 2019):

1. Explore the $\boldsymbol{\theta}$ space via the approximation $\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})$. Specifically, find the mode of $\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})$ and identify a set of points $\{\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^K\}$ in the areas of high density.

2. For these K points, compute $\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})$ using (15).

3. Calculate $\widehat{\pi}(x_i|\boldsymbol{\theta}^k, \boldsymbol{y})$ for $k = 1, \dots, K$ using one of Gaussian, Laplace, simplified Laplace.

4. Use numerical integration to approximate the marginal,

$$\widehat{\pi}(x_i|\boldsymbol{y}) = \sum_{k=1}^{K} \widehat{\pi}(x_i|\boldsymbol{\theta}^k, \boldsymbol{y}) \times \widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})\Delta_k, \tag{16}$$

   using points and weights $\{\boldsymbol{\theta}^k, \Delta_k, k = 1, \dots, K\}$.

Exploring the $\boldsymbol{\theta}$ space

First, a “good” parameterization is found (often this is achieved by simply transforming to the real line), we assume that $\boldsymbol{\theta}$ satisfies this; also let $\dim(\boldsymbol{\theta}) = m$.

Second, find the mode, $\boldsymbol{\theta}^\star$, and the Hessian matrix $\boldsymbol{H}$; let $\boldsymbol{H}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$ be the eigen decomposition, then form the new standardized variable:

$$\boldsymbol{z} = (\boldsymbol{V}\boldsymbol{\Lambda}^{1/2})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star),$$

which adjusts for location, scale, and rotation.

Exploring the $\boldsymbol{\theta}$ space

Rue *et al.* (2009) describe three methods for exploration:

1. grid: This approach builds a grid for the standardized variable $\boldsymbol{z}$. Unfortunately the number of points grows exponentially with m; if we use p points in each dimension, p^m are required in total.

2. empirical Bayes: just take the posterior mode only, i.e., a single point.

3. CCD: use a classical design, specifically the central composite design (CCD) – integration points are placed on spheres.

Grid versus CCD

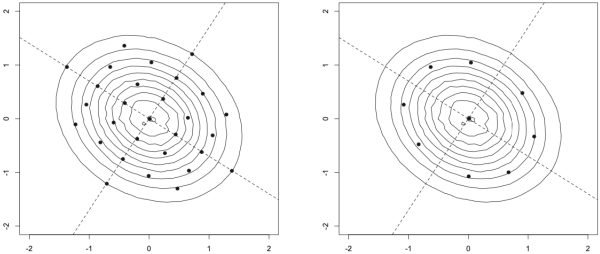


Figure 10: Grid (left) and CCD (right) points for numerical integration, from Wang *et al.* (2018).

INLA: Posterior sampling

Marginals are the standard output of INLA, but various operations may be carried out using the functions:

- `inla.dmarginal` for density values
- `inla.pmarginal` for the CDF
- `inla.qmarginal` for quantiles
- `inla.rmarginal` for random samples
- `inla.hpdmarginal` for HPD regions
- `inla.emarginal` computes the expected values of a function of a parameter
- `inla.tmarginal` calculates the marginal distribution of a transformation of a latent variable or hyperparameter.

INLA: Practical Advice

Some functionals cannot be obtained using these functions, so samples may be drawn from an approximation to the posterior[1], and manipulated:

- `inla.posterior.sample()` draws samples from the approximate posterior distribution of β and $\boldsymbol{\theta}$.
- To make use of this function, use `control.compute = list(config = TRUE)` in the INLA model fit.
- Included in the arguments is `selected` which allows only specific components to be sampled.
- In general, the returned sample contains

```
"hyperpar" "latent" "logdens"
```

[1] for the latent field $\boldsymbol{x}$ we sample from a mixture of multivariate Gaussians, where the weights correspond to the integration weights (for the grid and CCD options).

Pros and Cons of INLA

Advantages:

- Quite widely applicable: Generalized Linear Mixed Models (GLMMs) including temporal and spatial error terms – many book-length treatments now available, for example, Blangiardo and Cameletti (2015); Wang *et al.* (2018); Krainski *et al.* (2018).
- Very fast.
- An `R` package is available.

Disadvantages:

- Restricted to models with Gaussian random effects – Template Model Builder is more flexible, but the `TMB` package not have the same user-friendly interface as the `INLA` package.
- Experience required to assess when the approximation is failing, though lots of empirical evidence being gathered.

Markov chain Monte Carlo (MCMC)

MCMC has been around for a long time, being used by physicists, such as Metropolis *et al.* (1953).

Gelfand and Smith (1990) brought the technique to the general statistical world, and illustrated its power in a series of papers, beginning with Gelfand *et al.* (1990).

There are many implementations of MCMC in `R`, that can be used for generic modeling:

- `WinBUGS`,
- `OpenBUGS`,
- Just Another Gibbs Sampler: `JAGS`,
- `NIMBLE`,
- `Stan`.

Markov chain Monte Carlo (MCMC)

The fundamental idea behind MCMC is:

> To construct a Markov chain over the parameter space, with invariant distribution the posterior distribution of interest.

Specifically, consider a random variable $\boldsymbol{X}$ with support $\mathbb{R}^p$ and density $\pi(\cdot)$.

A sequence of random variables $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots$ is called a Markov chain on a state space $\mathbb{R}^p$ if for all t and for all measurable sets A:

$$\Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t-1)}, \ldots, \boldsymbol{X}^{(0)}\right) = \Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}\right)$$

so that the probability of moving to any set A at time $t+1$ only depends on where we are at time t.

Furthermore, for a homogeneous Markov chain:

$$\Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}\right) = \Pr\left(\boldsymbol{X}^{(1)} \in A \mid \boldsymbol{X}^{(0)}\right).$$

Markov chain Monte Carlo

If there exists $p(\boldsymbol{x}, \boldsymbol{y})$ such that

$$\Pr(\boldsymbol{X}_1 \in A \mid \boldsymbol{x}) = \int_A p(\boldsymbol{x}, \boldsymbol{y}) \, d\boldsymbol{y},$$

then $p(\boldsymbol{x}, \boldsymbol{y})$ is called the transition kernel density.

A probability distribution $\pi(\cdot)$ on $\mathbb{R}^p$ is called an invariant distribution of a Markov chain with transition kernel density $p(\boldsymbol{x}, \boldsymbol{y})$ if so-called global balance holds:

$$\pi(\boldsymbol{y}) = \int_{\mathbb{R}^p} \pi(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y}) \, d\boldsymbol{x}.$$

A Markov chain is called reversible if

$$\pi(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) p(\boldsymbol{y}, \boldsymbol{x}) \tag{17}$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{x} \neq \boldsymbol{y}$.

It can shown that if (17) holds then $\pi(\cdot)$ is the invariant distribution which is useful since (17) can be easy to check.

Markov chain Monte Carlo

A key idea is that if we have an invariant distribution then we can evaluate long term, or ergodic, averages of realizations of the chain, and these are estimates of the appropriate functions of $\pi(\cdot)$.

This is crucial for making inference in a Bayesian setting since it means we can estimate quantities of interest such as posterior means, medians, etc.

Only very mild conditions are typically required to ensure that $\pi(\cdot)$ is the invariant distribution, typically aperiodocity and irreducibility.

Markov chain Monte Carlo

A chain if periodic if there are places in the parameter space that can only be reached at certain regularly spaced times, otherwise it is aperiodic.

A Markov chain with invariant distribution $\pi(\cdot)$, is *irreducible* if for any starting point there is positive probability of entering any set to which $\pi(\cdot)$ assigns positive probability.

Suppose that $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ represents the sample path of the Markov chain.

Then expectations with respect to the invariant distribution

$$\mu = \mathrm{E}[g(\boldsymbol{x})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\ d\boldsymbol{x}$$

may be approximated by $\widehat{\mu}_m = \frac{1}{m}\sum_{t=1}^{m} g(\boldsymbol{x}^{(t)})$.

Monte Carlo standard errors are more difficult to obtain than in the independent sampling case.

Markov chain Monte Carlo

The Markov chain law of large numbers (the ergodic theorem) tells us that

$$\widehat{\mu}_m \rightarrow_{a.s.} \mu$$

as $m \rightarrow \infty$ and the Markov chain central limit theorem states that

$$\sqrt{m}(\widehat{\mu}_m - \mu) \rightarrow_d \mathrm{N}(0, \tau^2)$$

where

$$\tau^2 = \mathrm{var}\left[g(\boldsymbol{x}^{(t)})\right] + 2\sum_{k=1}^{\infty} \mathrm{cov}\left[g(\boldsymbol{x}^{(t)}), g(\boldsymbol{x}^{(t+k)})\right] \tag{18}$$

and the summation term accounts for the dependence in the chain.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm provides a very flexible method for defining a Markov chain.

At iteration t of the Markov chain's evolution suppose the current point is $\boldsymbol{x}^{(t)}$.

The following steps provide the new point $\boldsymbol{x}^{(t+1)}$:

1. Sample a point $\boldsymbol{y}$ from a proposal distribution $q(\cdot \mid \boldsymbol{x}^{(t)})$.
2. Calculate the acceptance probability:

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left[\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(t)})} \times \frac{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{y})}{q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)})}, 1 \right]. \tag{19}$$

3. Set

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{y} & \text{with probability } \alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) \\ \boldsymbol{x}^{(t)} & \text{otherwise.} \end{cases}$$

The Metropolis-Hastings Algorithm

In a Bayesian context, the term,

$$\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}$$

in equation (19) is the ratio of the posterior density at the proposed point ***y*** to the current point ***x***.

Since we are taking the ratio the normalizing constants in the posterior cancel, which is crucial since these are typically unavailable.

The Metropolis-Hastings Algorithm

The second term in (19) is the ratio of the density of moving from

$$\boldsymbol{y} \rightarrow \boldsymbol{x}^{(t)}$$

to the density of moving from

$$\boldsymbol{x}^{(t)} \rightarrow \boldsymbol{y}$$

and it is this term that guarantees global balance and hence that the Markov chain has the correct invariant distribution.

In an independence chain the proposal distribution does not depend on the current point, i.e. $q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)})$ is independent of $\boldsymbol{x}^{(t)}$.

We now consider a special case of the algorithm which is particularly easy to implement and is widely used.

The Metropolis Algorithm

Suppose the proposal distribution is symmetric in the sense that

$$g(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = g(\boldsymbol{x}^{(t)} \mid \boldsymbol{y}).$$

In this case the product of ratios in (19) simplifies to

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min \left[\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}, 1 \right]$$

so that only the ratio of target densities is required.

In the random walk Metropolis algorithm $q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = q(\ |\boldsymbol{y} - \boldsymbol{x}^{(t)}|\ )$, with common choices for $q(\cdot)$ being normal or uniform distributions.

In a range of circumstances an acceptance probability of around 30% is optimal, which may be obtained by tuning the proposal density, the variance in a normal proposal, for example.

The balancing act is between having high acceptance rates with small movement or having low acceptance rates with large movement.

The Gibbs Sampler

We describe a particularly popular algorithm for simulating from a Markov chain, the Gibbs sampler.

We describe two flavors: the sequential Gibbs sampler and the random scan Gibbs sampler.

In the following, let $\boldsymbol{x}_{-i}$ represent the vector $\boldsymbol{x}$ with the i-th variable removed, i.e. $\boldsymbol{x}_{-i} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p]$.

The Gibbs Sampler

The sequential scan Gibbs sampling algorithm starts with some initial value $\boldsymbol{x}^{(0)}$ and then, with current point

$$\boldsymbol{x}^{(t)} = [x_1^{(t)}, \ldots, x_p^{(t)}],$$

undertakes the following p steps to produce a new point

$$\boldsymbol{x}^{(t+1)} = [x_1^{(t+1)}, \ldots, x_p^{(t+1)}]$$

- Sample $x_1^{(t+1)} \sim \pi_1 \left(x_1 \mid \boldsymbol{x}_{-1}^{(t)} \right)$
- Sample $x_2^{(t+1)} \sim \pi_2 \left(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)} \right)$

  $\vdots$
- Sample $x_p^{(t+1)} \sim \pi_p \left(x_p \mid \boldsymbol{x}_{-p}^{(t+1)} \right)$.

The Gibbs Sampler

The beauty of the Gibbs sampler is that the often hard problem of sampling for the full p-dimensional variable $\boldsymbol{x}$ has been broken into sampling for each of the p variables in turn via the conditional distributions.

We now illustrate that the Gibbs sampling algorithm produces a transition kernel density that gives the required stationary distribution.

We do this by showing that each component is a Metropolis-Hastings step.

Consider a single component move in the Gibbs sampler from the current point $\boldsymbol{x}^{\text{cur}}$ to the new point $\boldsymbol{x}^{\text{new}}$, with $\boldsymbol{x}^{\text{new}}$ obtained by replacing the i-th component in $\boldsymbol{x}^{\text{cur}}$ with a draw from the full conditional $\pi\left(x_i \mid \boldsymbol{x}_{-i}^{\text{cur}}\right)$.

We view this move in light of the Metropolis-Hastings algorithm in which the proposal density is the full conditional itself.

The Gibbs Sampler

Then the Metropolis-Hastings acceptance ratio becomes

$$\begin{aligned}\alpha(\boldsymbol{x}^{\text{cur}}, \boldsymbol{x}^{\text{new}}) &= \min\left[\frac{\pi\left(x_i^{\text{new}}, \boldsymbol{x}_{-i}^{\text{cur}}\right)\pi\left(x_i^{\text{cur}} \mid \boldsymbol{x}_{-i}^{\text{cur}}\right)}{\pi\left(x_i^{\text{cur}}, \boldsymbol{x}_{-i}^{\text{cur}}\right)\pi\left(x_i^{\text{new}} \mid \boldsymbol{x}_{-i}^{\text{cur}}\right)}, 1 \right] \\ &= \min\left[\frac{\pi\left(\boldsymbol{x}_{-i}^{\text{cur}}\right)}{\pi\left(\boldsymbol{x}_{-i}^{\text{cur}}\right)}, 1 \right] = 1\end{aligned}$$

because

$$\pi\left(\boldsymbol{x}_{-i}^{\text{cur}}\right) = \pi\left(x_i^{\star}, \boldsymbol{x}_{-i}^{\text{cur}}\right) / \pi\left(x_i^{\star} \mid \boldsymbol{x}_{-i}^{\text{cur}}\right).$$

Consequently, when we use full conditionals as our proposals in the Metropolis-Hastings step we always accept.

The Gibbs Sampler

This means that drawing from a full conditional distribution produces a Markov chain with stationary distribution $\pi(\boldsymbol{x})$.

Clearly, we cannot keep updating only the *i*-th component, because we will not be able to explore the whole state space this way, i.e. we do not have an irreducible Markov chain.

Therefore, we can update each component in turn, though this is not the only way to execute Gibbs sampling (though it is the easiest to implement and the most common).

The Gibbs Sampler

We can also randomly select an component to update.

This is called random scan Gibbs sampling.

- Sample a component i by drawing a random variable with probability mass function $[\alpha_1, \ldots, \alpha_p]$ where $\alpha_i > 0$ and $\sum_{i=1}^{p} \alpha_i = 1$.
- Sample $x_i^{(t+1)} \sim \pi_i \left(x_i \mid \boldsymbol{x}_{-i}^{(t)} \right)$.

In many cases, conjugacy can be exploited to derive the conditional distributions.

It is also common for sampling from a full conditional distribution to not require knowledge of the normalizing constant of the target distribution.

Combining Markov Kernels: Hybrid Schemes

Suppose we can construct *m* transition kernels, each with invariant distribution $\pi(\cdot)$. There are two simple ways to combine these transition kernels.

First, we can construct a Markov chain, where at each step we sequentially generate new states from all kernels in a predetermined order.

As long the new Markov chain is irreducible, then it will have the required invariant distribution and we can, for example, use the ergodic theorem on the samples from the new Markov chain.

Hence, we can combine Gibbs and Metropolis-Hastings steps.

One popular form is Metropolis within Gibbs in which all conditionals are sampled with Gibbs steps for the recognizable conditionals and Metropolis-Hastings for the remainder.

Combining Markov Kernels: Hybrid Schemes

In the second method of combining Markov kernels, we first create a probability vector $[\alpha_1, \ldots, \alpha_m]$, then randomly select kernel *i* with probability α_i and then use this kernel to move the Markov chain.

In general, one can be creative in the construction of a Markov chain, but care must be taken to ensure the proposed chain is “legal”, in the sense of having the required stationary distribution.

As an example, a chain with a Metropolis step that keeps proposing points until the *k*-th point, with $k \geqslant 1$, is accepted does not have the correct invariant distribution.

Implementation Details

In practice, there are a number of important issues that require thought when implementing MCMC.

A crucial question is how large m should be chosen in order to obtain a reliable Monte Carlo estimate.

The Markov chain will display better mixing properties if the parameters are approximately independent in the posterior.

Implementation Details

In an extreme case, if we have independence then

$$\pi(x_1, \ldots, x_p) = \prod_{i=1}^{p} \pi(x_i)$$

and Gibbs sampling via the conditional distributions $\pi(x_i), i = 1, \ldots, p$, equates to direct sampling from the posterior.

Dependence in the Markov chain may be greatly reduced by sampling simultaneously for variables that are highly depend, a strategy known as blocking.

Reparameterization

Reparameterization may also be helpful in this regard.

As the blocks become larger the acceptance rate may be reduced to an unacceptably low level in which case there is a trade-off in the size of blocks to use.

Some chains may be very slow mixing and an examination of autocorrelation aids in deciding on the number of iterations required.

If storage of samples is an issue then one may decide to thin the chain by only collecting samples at equally spaced intervals.

Reparameterization

A number of methods have been proposed for diagnosing convergence.

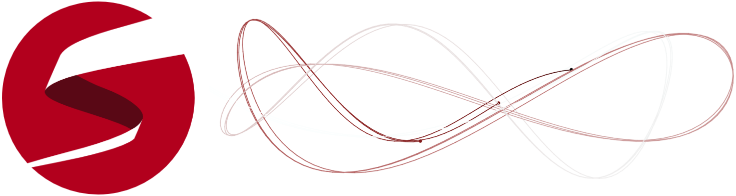
Trace plots provide a useful method for detecting problems with MCMC convergence and mixing. Ideally, trace plots of unnormalized log posterior and model parameters should look like stationary time series.

Slowly mixing Markov chains produce trace plots with high autocorrelation, which can be further visualized by autocorrelation plots at different lags.

Slow mixing does not imply lack of convergence, however, but that more samples will be required for accurate inference (as can be seen from (18).

When examining trace plots and autocorrelations it is clearer to work with parameters transformed to $\mathbb{R}$.

Stan



- Stan is similar to WinBUGS, JAGS, but new and improved.
- Coded in C++, for faster updating, it runs the *No U-Turn Sampler* (Hoffman and Gelman, 2014) – cleverer than WinBUGS' algorithms (it uses gradient information on the posterior).
- The rstan package lets you run chains from R (just like R2WinBUGS).
- Some modeling limitations – no discrete parameters – but becoming very popular; works well with some models where WinBUGS would struggle
- Basically the same modeling language as WinBUGS – but Stan allows R-style vectorization
- Requires declarations (like C++) – unlike WinBUGS, or R – so models require a bit more typing...

Example: Lung Cancer and Radon

For direct comparison with methods applied in the frequentist chapter, we assume an improper flat prior on $\boldsymbol{\beta} = [\beta_0, \beta_1]$ so that the posterior $p(\boldsymbol{\beta} \mid \boldsymbol{y})$ is proportional to the likelihood.

We begin by implementing a Metropolis random walk algorithm based on a pair of univariate normal distributions.

In this example the Gibbs sampler is less appealing since the required conditional distributions do not assume known forms.

The first step is to initialize $\beta_0^{(0)} = \widehat{\beta}_j$, where $\widehat{\beta}_j$, $j = 0, 1$, are the MLEs
.

Example: Lung Cancer and Radon

We then iterate, at iteration t, between:

1. Generate $\beta_0^\star \sim \mathrm{N}(\beta_0^{(t)}, c_0 \widehat{V}_0)$, where $\widehat{V}_0$ is the asymptotic variance of $\widehat{\beta}_0$. Calculate the acceptance probability:

$$\alpha_0(\beta_0^\star, \beta_0^{(t)}) = \min\left[\frac{p(\beta_0^\star, \beta_1^{(t)} \mid \boldsymbol{y})}{p(\beta_0^{(t)}, \beta_1^{(t)} \mid \boldsymbol{y})}, 1 \right]$$

and set

$$\beta_0^{(t+1)} = \begin{cases} \beta_0^\star & \text{with probability } \alpha_0(\beta_0^\star, \beta_0^{(t)}), \\ \beta_0^{(t)} & \text{otherwise.} \end{cases}$$

2. Generate $\beta_1^\star \sim \mathrm{N}(\beta_1^{(t)}, c_1 \widehat{V}_1)$, where $\widehat{V}_1$ is the asymptotic variance of $\widehat{\beta}_1$. Calculate the acceptance probability:

$$\alpha_1(\beta_1^\star, \beta_1^{(t)}) = \min\left[\frac{p(\beta_0^{(t+1)}, \beta_1^\star \mid \boldsymbol{y})}{p(\beta_0^{(t+1)}, \beta_1^{(t)} \mid \boldsymbol{y})}, 1 \right]$$

and set

$$\beta_1^{(t+1)} = \begin{cases} \beta_1^\star & \text{with probability } \alpha_1(\beta_1^\star, \beta_1^{(t)}), \\ \beta_1^{(t)} & \text{otherwise.} \end{cases}$$

Example: Lung Cancer and Radon

The constants c_0 and c_1 are chosen to provide a trade-off between gaining a high proportion of acceptances, and moving around the support of the parameter space.

This is illustrated in Figure 11 where the realized parameters from the first 1000 iterations of two Markov chains are plotted.

In panels (a) and (d) we chose $c_0 = c_1 = c = 0.1$ and in panels (b) and (e) $c_0 = c_1 = c = 2$.

For $c = 0.1$ the acceptance rate is 0.90 but movement around the space is slow, as indicated by the meandering nature of the chain, while for $c = 2$ the moves tend to be larger but the chain sticks at certain values, as seen by as horizontal runs of points, with an acceptance rate of 0.14.

Example: Lung Cancer and Radon

Figure 13(a) shows a scatterplot representation of the joint distribution $p(\beta_0, \beta_1 \mid \boldsymbol{y})$.

We clearly see the strong negative dependence; the asymptotic correlation between the MLEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is -0.90, and the posterior correlation between β_0 and β_1 is -0.90 also (the correspondence between these correlations is not surprising since the sample size is large and the prior is flat).

The strong negative dependence is evident in each of the first two columns.

Example: Lung Cancer and Radon

Figure 12 shows the autocorrelations between sampled parameters at lags of between 1 and 40. The top row is for β_0 and the bottom is for β_1.

In panels (a) and (d) the autocorrelations are high because of the small movements of the chain.

The dependence in the chain may be reduced via reparameterization, or by generation from a bivariate proposal.

We implement the latter with variance-covariance matrix equal to $c \times \text{var}(\widehat{\boldsymbol{\beta}})$. The acceptance rate for the bivariate proposal with $c = 2$ is 0.29, which is reasonable.

Example: Lung Cancer and Radon

We then iterate the following:

1. Generate $\boldsymbol{\beta}^\star \sim \mathsf{N}_2(\boldsymbol{\beta}^{(t)}, c\widehat{\boldsymbol{V}})$, where $\widehat{\boldsymbol{V}}$ is the asymptotic variance of $\widehat{\boldsymbol{\beta}}$.
2. Calculate the acceptance probability

$$\alpha(\boldsymbol{\beta}^\star, \boldsymbol{\beta}^{(t)}) = \min\left[\frac{p(\boldsymbol{\beta}^\star \mid \boldsymbol{y})}{p(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{y})}, 1\right]$$

and set

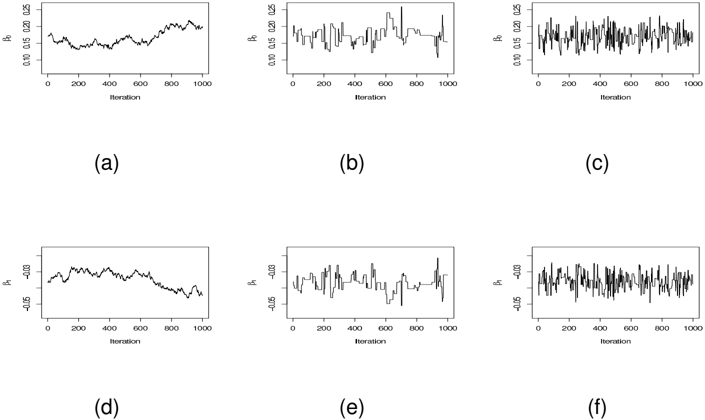
$$\boldsymbol{\beta}^{(t+1)} = \begin{cases} \boldsymbol{\beta}^\star & \text{with probability } \alpha(\boldsymbol{\beta}^\star, \boldsymbol{\beta}^{(t)}), \\ \boldsymbol{\beta}^{(t)} & \text{otherwise.} \end{cases}$$

Example: Lung Cancer and Radon

Note that the choice of c and the dependence in the chain do not jeopardize the invariant distribution, but rather the length of chain until practical convergence is reached and the number of points required for summarization.

More points are required when there is positive dependence in successive iterates, which is clear from (18).

The final column of Figure 11 shows the sample path from the bivariate proposal, with good movement and no dependence between the parameters.



(a) (b) (c)

(d) (e) (f)

FIGURE 11: Sample paths from Metropolis-Hastings algorithms for β_0 (top row) and β_1 (bottom row) for the lung cancer and radon data. In the left column the proposal random walk has small variance; in the center column large variance, and in the right column we use a bivariate proposal.

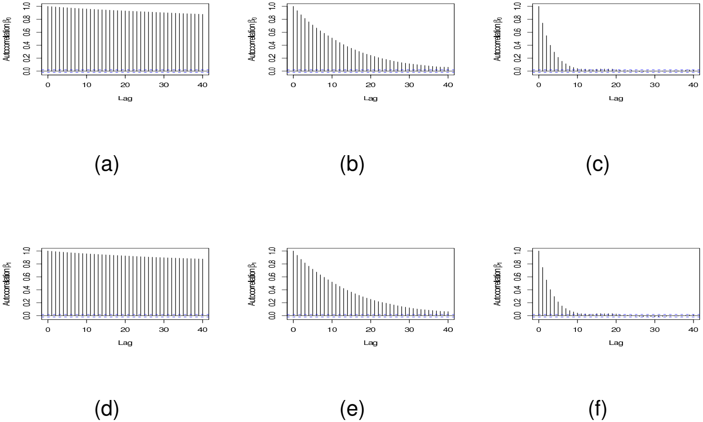


FIGURE 12: Autocorrelation functions for β_0 (top row) and β_1 (bottom row) for the lung cancer and radon data. First column: univariate random walk, $c = 0.1$, second column: univariate random walk, $c = 2$, third column: bivariate random walk, $c = 2$.

Example: Lung Cancer and Radon

Figure 13 shows inference for the reparameterized model

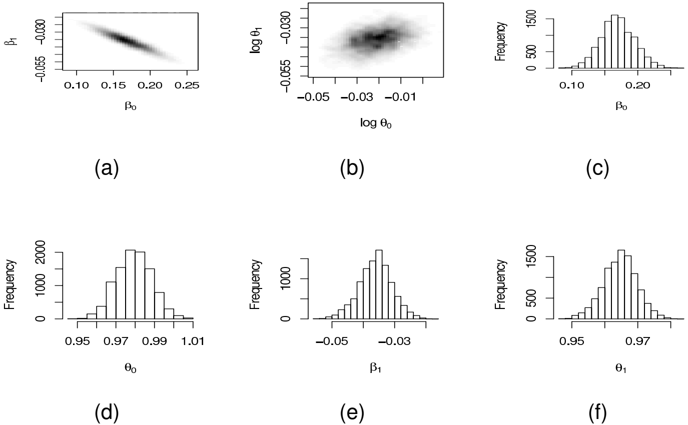
$$Y_i \mid \boldsymbol{\theta} \sim_{ind} \text{Poisson}(E_i \theta_0 \theta_1^{x_i - \overline{x}})$$

where $\theta_0 = \exp(\beta_0 + \beta_1 \overline{x}) > 0$ and $\theta_1 = \exp(\beta_1) > 0$.

Figure 13(d) shows the bivariate posterior for $\log \theta_0, \log \theta_1$ and demonstrates that the parameters are virtually independent (the correlation is -0.03).

Panels (e) and (f) show histogram representations of the posteriors of interest $p(\theta_0 \mid \boldsymbol{y})$ and $p(\theta_1 \mid \boldsymbol{y})$.

The posterior median (95% credible interval) for $\exp(\beta_1)$ is $0.965\ [0.954, 0.975]$ which is almost identical to the asymptotic inference under a Poisson model, which is again not surprising given the large sample size.



(a) (b) (c)

(d) (e) (f)

FIGURE 13: Posterior summaries for the lung cancer and radon data: (a) $p(\beta_0, \beta_1 \mid \boldsymbol{y})$, (b) $p(\log\ \theta_0, \log\ \theta_1 \mid \boldsymbol{y})$, (c) $p(\beta_0 \mid \boldsymbol{y})$, (d) $p(\theta_0 \mid \boldsymbol{y})$, (e) $p(\beta_1 \mid \boldsymbol{y})$, (f) $p(\theta_1 \mid \boldsymbol{y})$.

Bayes Factors

Bayes Factors for Hypothesis Testing

- The Bayes factor provides a summary of the evidence for a particular hypothesis (model) as compared to another.
- The Bayes factor is

$$\text{BF} = \frac{\Pr(y|H_0)}{\Pr(y|H_1)}$$

  and so is simply the probability of the data under H_0 divided by the probability of the data under H_1.
- Values of BF > 1 favor H_0 while values of BF < 1 favor H_1.
- Note the similarity to the likelihood ratio

$$\text{LR} = \frac{\Pr(y|H_0)}{\Pr(y|\widehat{\theta})}$$

  where $\widehat{\theta}$ is the MLE under H_1.
- If there are no unknown parameters in H_0 and H_1 (for example, in a binomial with parameter θ, $H_0 : \theta = 0.5$ versus $H_1 : \theta = 0.3$), then the Bayes factor is identical to the likelihood ratio.

Calibration of Bayes Factors

- Kass and Raftery (1995) suggest intervals of Bayes factors for reporting:

1/Bayes Factor	Evidence Against H_0
1 to 3.2	Not worth more than a bare mention
3.2 to 20	Positive
20 to 150	Strong
$>$150	Very strong

- These provide a guideline, but should not be followed without question.

Model selection

Suppose we wish to choose between models M_0 and M_1.

We may calculate the posterior odds

$$\frac{p(M_1|\boldsymbol{y})}{p(M_0|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_0)} \times \frac{p(M_1)}{p(M_0)},$$

where

$$\frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_0)}$$

is the Bayes factor and

$$\frac{p(M_1)}{p(M_0)},$$

the prior odds.

The posterior odds is highly sensitive to the prior odds.

Model selection

In particular in conjugate cases, it is easier to evaluate posterior and predictive distributions since they must integrate to one.

Calculating the normalizing constant is more troublesome since we must keep track of all the constants.

If Y is discrete then $0 < p(\boldsymbol{y}|M) \leqslant 1$, for y continuous, $p(\boldsymbol{y}) > 0$ only, in particular this density value may be very small or very big...

The Bayes factor may be thought of as the Bayesian version of the likelihood ratio statistic.

Hypothesis Testing

Suppose we are interested in choosing between the two models $M_0 : \theta = 0$ and $M_1 : \theta \neq 0$.

We assume the prior is given by $\theta \sim \mathrm{N}(0, v)$ where $v = \sigma^2/k$.

The Bayes factor of M_0 versus M_1 is,

$$\begin{aligned} B_{01} &= \frac{p(\boldsymbol{y}|M_0)}{p(\boldsymbol{y}|M_1)} \\ &= \frac{p(\boldsymbol{y}|M_0)}{\int p(\boldsymbol{y}|\theta, M_1)p(\theta|M_1)d\theta} \\ &= \frac{\cancel{(2\pi\sigma^2)^{-n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2\right]}{\cancel{(2\pi\sigma^2)^{-n/2}}(2\pi)^{-1/2}(\sigma^2/k)^{-1/2}\int \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\theta)^2 - \frac{k\theta^2}{2\sigma^2}\right]d\theta} \\ &= \left(\frac{k+n}{k}\right)^{1/2}\exp\left[-\frac{n\bar{y}^2}{2\sigma^2}\times\frac{n}{k+n}\right]. \end{aligned}$$

Hypothesis Testing

Intuitively, we would expect M_1 to be favored as the prior variance increases.

However, as $k \to 0$ (so $p(\theta = 0 | M_1)$ is decreasing) we note that $B_{01} \to (1 + n/k)^{1/2} \to \infty$, that is, we would conclude that M_0 is favored.

This behavior is known as Lindley's paradox: the numerator is an ordinate on the density while the denominator is the likelihood of the data, integrated over the prior – when the latter is flat there are a large number of likelihood contributions that are essentially zero (and these are not wiped out by the prior).

(In estimation, roughly speaking, one can sometimes get away with flat priors since the arbitrary constant cancels in numerator and denominator of Bayes theorem).

Message: for model selection via Bayes factors we must be careful with prior specification.

Example: Bayes Factors for Binomial Data

Consider a binomial experiment in which are interested in $H_0 : \theta = 0.5$ versus $H_1 : \theta \neq 0.5$.

The numerator and denominator of the Bayes factor are:

$$\begin{aligned}
\Pr(y|H_0) &= \binom{N}{y} 0.5^y 0.5^{N-y} \\
\Pr(y|H_1) &= \int_0^1 \binom{N}{y} \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \\
&= \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}
\end{aligned}$$

We have already seen the denominator calculation, when we normalized the posterior.

Values Taken by the Negative Log Bayes Factor, as a Function of y

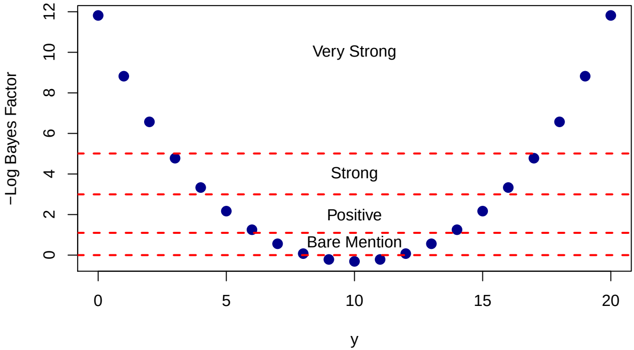


FIGURE 14: Negative Log Bayes factor as a function of $y|\theta \sim \text{Binomial}(20, \theta)$ for $y = 0, 1, \ldots, 20$ and $a = b = 1$. High values indicate evidence against the null.

Bayesian Model Averaging

If a discrete number of models is considered then model averaging provides an alternative means of assessing model uncertainty – very different course of action then frequentist approaches that look for estimators that are valid under model misspecification.

The Bayesian machinery handles multiple models in a very straightforward fashion since essentially the unknown model is treated as an additional discrete parameter.

Let $M_1, \ldots, M_J$ denote the J models under consideration, and $\boldsymbol{\theta}_j$ the parameters of the j-th model.

Suppose, for illustration, there is a parameter of interest ϕ (which we assume is univariate) that is well-defined for each of the J models under consideration.

Bayesian Model Averaging

The posterior for ϕ is a mixture over the J individual model posteriors:

$$p(\phi \mid \boldsymbol{y}) = \sum_{j=1}^{J} p(\phi \mid M_j, \boldsymbol{y}) \times \Pr(M_j \mid \boldsymbol{y})$$

where

$$\begin{aligned}
p(\phi \mid M_j, \boldsymbol{y}) &= \int p(\phi \mid \boldsymbol{\theta}_j, M_j, \boldsymbol{y}) p(\boldsymbol{\theta}_j \mid M_j, \boldsymbol{y}) \, d\boldsymbol{\theta}_j \\
&= \frac{1}{p(\boldsymbol{y} \mid M_j)} \int p(\phi \mid \boldsymbol{\theta}_j, M_j, \boldsymbol{y}) p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j \mid M_j) \, d\boldsymbol{\theta}_j, \\
\Pr(M_j \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid M_j) \Pr(M_j)}{p(\boldsymbol{y})} \\
&= \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j \mid M_j) \, d\boldsymbol{\theta}_j \Pr(M_j)}{p(\boldsymbol{y})}
\end{aligned}$$

and with $\Pr(M_j)$ the prior belief in model j and $p(\boldsymbol{\theta}_j \mid M_j)$ the prior on the parameters of model M_j.

Bayesian Model Averaging

The marginal probabilities of the data under the different models are calculated as

$$p(\boldsymbol{y} \mid M_j) \quad = \quad \int p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j \mid M_j) \, d\boldsymbol{\theta}_j,$$

with

$$p(\boldsymbol{y}) \quad = \quad \sum_{j=1}^{J} p(\boldsymbol{y} \mid M_j) \, \Pr(M_j)$$

To summarize the posterior for ϕ, we can examine the posterior marginal distribution.

Bayesian Model Averaging

We might summarize via the posterior mean

$$\mathsf{E}[\phi \mid \boldsymbol{y}] = \sum_{j=1}^{J} \mathsf{E}[\phi \mid \boldsymbol{y}, M_j] \times \Pr(M_j \mid \boldsymbol{y}),$$

which is simply the average of the posterior means across models, weighted by the posterior weight received by each model.

The posterior variance is

$$\begin{aligned} \mathrm{var}(\phi \mid \boldsymbol{y}) &= \sum_{j=1}^{J} \mathrm{var}(\phi \mid \boldsymbol{y}, M_j) \times \Pr(M_j \mid \boldsymbol{y}) \\ &+ \sum_{j=1}^{J} \left\{ \mathsf{E}[\phi \mid \boldsymbol{y}, M_j] - \mathsf{E}[\phi \mid \boldsymbol{y}] \right\}^2 \times \Pr(M_j \mid \boldsymbol{y}) \end{aligned}$$

which averages the posterior variances concerning ϕ in each model, with the addition of a term that accounts for between-model differences in the mean.

Bayesian Model Averaging

- Performing model averaging over models which represent different scientific theories is also not appealing if the search for a causal explanation is sought.

- Parameter interpretation can be tricky when BMA is used.

- If prediction is the aim then model averaging is much more appealing since parameter interpretation is often irrelevant – ensembles of model methods are popular and successful in prediction settings.

- A disadvantage of model averaging is that it may encourage the user to believe they have accounted for all model uncertainty, which is a dangerous conclusion to draw.

Discussion

- Penalized complexity (PC) priors are a recent class (Simpson *et al.*, 2017) with very appealing characteristics.

- Never forget that Bayesian summaries are conditional on all model assumptions, including appropriateness of prior.

- Bayesian computation has come a long way, and `INLA` and `Stan` are very convenient (and continually being improved), once one gets up to speed.

References

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra, R., Tedrow, J., *et al.* (2020). Covid-19 antibody seroprevalence in Santa Clara county, California. *MedRxiv*.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley and Sons.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990). Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–985.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society, Series A*. To appear.

Hoffman, M. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Liang, K.-Y. and McCullagh, P. (1993). Case studies in binary dispersion. *Biometrics*, **49**, 623–630.

Martino, S. and Riebler, A. (2019). Integrated nested laplace approximations (INLA). *arXiv preprint arXiv:1907.01248*.

Metropolis, N., Rosenbluth, A., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

Savage, S. A., Gerstenblith, M. R., Goldstein, A. M., Mirabello, L., Fargnoli, M. C., Peris, K., and Landi, M. T. (2008). Nucleotide diversity and population differentiation of the melanocortin 1 receptor gene, mc1r. *BMC Genetics*, **9**, 31.

Simpson, D., Rue, H., Riebler, A., Martins, T., and Sørbye, S. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, **32**, 1–28.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Wang, X., Yue, Y., and Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.