A simple example of Bayesian statistics and MCMC estimation

Michael Ash[*]

16 October 2024

The goal is to estimate the parameters (mean μ and variance σ^2) of normally distributed data,

$$y_i \sim N(\mu, \sigma^2), \ i = 1, \ldots, n$$

or more accurately to estimate the posterior distribution of the parameters.

The mean is μ, but instead of σ^2 we will examine h, the inverse of the variance, i.e., $h = 1/\sigma^2$. Hence the actual target is

$$y_i \sim N(\mu, h^{-1}), \ i = 1, \ldots, n.$$

The reason for this construction is that it's more convenient to use a Gamma distribution for the distribution of h rather than an inverted- (or inverse-) Gamma distribution for the distribution of σ^2.

One of the key pieces of Bayesian statistics is the prior distribution, subjective belief about the distribution of the parameters before we examine the data. There are two parameters, μ and h, for which we must have prior beliefs.

The prior for μ is

$$\mu \sim N(\mu_0, h_0^{-1}).$$

Thus, our belief about μ is that it is normally distributed with mean μ_0 and variance h_0^{-1}. The density function for this prior, but for some proportionality terms, is $\pi(\mu) \propto \exp\left[-\frac{h_0}{2}(\mu - \mu_0)^2\right]$. Note that this is a rather diffuse prior distribution because the guess about the spread of our beliefs about μ, the mean of y, is the same as the σ^2 variance of the underlying distribution of y_i itself (not, for example, the CLT-implied σ^2/n which is a much tighter distribution—note that we don't yet have data, including n).[1]

The prior for h is

$$h \sim G(\alpha_0/2, \delta_0/2),$$

where $G(\cdot, \cdot)$ is the Gamma distribution. The Gamma distribution depends on two parameters ("hyperparameters" with respect to the main model), in this

case, α and δ with initial ("prior") values of α_0 and δ_0. Although the Gamma distribution is somewhat obscure, it's probably easiest to think about its parameters by noting a special case of the Gamma distribution: $G(\nu/2, 1/2)$ is the same as a chi-squared distribution χ^2_ν. First, this connects the reader to the inverted chi-squared distribution of estimates of the variance in frequentist statistics. As with the chi-squared distribution and as to be expected with variance, the density is only non-zero for positive arguments, i.e., the variance must be positive. Second the special case gives some sense of the magnitude of the metaparameters, α around twice the sample size and δ around twice the sum of squared errors are plausible values (note that both terms get halved as arguments in $G(\cdot, \cdot)$.[2] The density for this prior, again but for some proportionality terms, is $\pi(h) \propto h^{\alpha_0/2-1} \exp\left[-\frac{\delta_0 h}{2}\right]$.

It's worth pointing out at this point that the prior μ_0 for μ depends on the prior h_0 for h but not vice versa. The prior density for both that we work with will really be

$$\pi(\mu, h) = \pi(\mu|h)\pi(h).$$

The object of interest in Bayesian statistics is the posterior distribution, which unites the prior with the data to produce updated beliefs about the distribution of the parameters. We've discussed the priors above. The data are represented by the likelihood function, which expresses how likely are these data if the parameters were given for the random process that generated the data, $f(y|\mu, h)$. (How likely are the data y for given values of the parameters of a normal distribution, μ and h?)

Note that the posterior, the object of interest, is how likely are (each of the possible values of) the parameters with the data given, or $\pi(\mu, h|y)$. Bayes formula is what allows us to compute the posterior, the object of interest, given the data and the priors:

$$\pi(\mu, h|y) \propto f(y|\mu, h) \times \pi(\mu, h).$$

It's helpful to read this out loud: After seeing the data, how likely is a particular value of μ? It depends on how likely that value of μ was initially considered to be times how likely the observed values of the data were to occur if that value of μ were in fact the case.

Bayes formula also demonstrates why we need priors; it would otherwise be impossible to convert the likelihood function, where the probability of the *data given parameters*, is the output, into the posterior distribution, where the probability of *parameters given data*, is the output.

By the way, the formula uses "$\propto$" ("is proportional to") rather than "=" because a normalization term is dropped from the denominator of Bayes's formula.[3]

---

[2]The parameters are $G(\text{shape}, \text{rate})$ with "rate" sometimes described as "inverse scale." In many contexts, shape and rate are denoted α and β, which we'll avoid outside this footnote because α is otherwise used. For $X \sim G(\alpha, \beta)$, $E(X) = \alpha/\beta$ and $V(x) = \alpha/\beta^2$. In terms of the α and δ in the main body of the paper, $E(h) = \alpha/\delta$ and $V(h) = \alpha/(\delta^2/4)$. I made up guesses for the priors on h by checking that these terms gave plausible values.

[3]"But for a proportionality term" appears all over the Bayesian literature. The actual

Since we know that the data were generated from a normal distribution, we know the functional form of the likelihood function, which, again, expresses how likely is this particular outcome $(y_1, \ldots, y_n)$ with given parameters μ and h. In this case,

$$f(y|\mu, h) = h^{n/2} \exp\left[-\frac{h}{2}\sum(y_i - \mu)^2\right].$$

We have already specified the priors for the two parameters. We can multiply the likelihood by the priors to derive the posterior:

$$\pi(\mu, h|y) \propto \overbrace{h^{n/2} \exp\left[-\frac{h}{2}\sum(y_i - \mu)^2\right]}^{\text{Likelihood function}} \times \overbrace{\exp\left[-\frac{h_0}{2}(\mu - \mu_0)^2\right]}^{\text{Prior density of } \mu} \times \overbrace{h^{\alpha_0/2-1}\exp\left[-\frac{\delta_0 h}{2}\right]}^{\text{Prior density of } h}.$$

The first term on the right-hand side is the likelihood function and it is the only place where the data, the y_i, appear. The second term is the prior normal distribution of μ and the third term is the prior Gamma distribution of h. The analysis takes advantage of the dependence of μ on h and of the independence of h; so the prior terms can be simply multiplied together.

At this point, we additively combine the exponental terms, rearrange, add like terms, notice some instances of the mean of the data ($\bar{y} = \sum y_i/n$ is the MLE of μ), etc.[4] The rearranged equation can be elegantly split into conditional posterior distributions of h and μ.

The conditional posterior density of h is

$$\pi(h|\mu, y) \propto h^{(\alpha_0+n)/2-1}\exp\left[-h\frac{\delta_0 + \sum(y_i - \mu)^2}{2}\right],$$

which is the density function of $G[(\alpha_0+n)/2, (\delta_0+\sum(y_i-\mu)^2)/2]$. Thus, we have updated parameters for the Gamma distribution which describe the posterior Gamma distribution of h.

The old, prior first parameter was $\alpha_0/2$. The new, data-incorporating first parameter is $(\alpha_0 + n)/2$. The old, prior second parameter was $\delta_0/2$. The new data-incorporating second parameter is $(\delta_0 + \sum(y_i - \mu)^2)/2$

---

expression is $\pi(\mu, h|y) = \frac{f(y|\mu,h)\times\pi(\mu,h)}{f(y)}$. The $f(y)$ in the denominator is necessitated by Bayes's formula $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$, i.e., $f(y)$ is the $P(B)$ term, which is needed to normalize the expression for $P(A|B) = P(B|A)P(A)/P(B)$ to be between 0 and 1. It is ok to replace "=" with "$\propto$" thereby "ignoring" $f(y)$ because the marginal distribution of the data does not vary and, hence, $1/f(y)$ simply scales the density distribution up or down without altering the location of peaks or valleys or areas of portions relative to the area of the whole. This replacement of "=" with "$\propto$" is particularly useful with continuously distributed parameters where relative probabilities are of interest. If the question involves parameters for a discrete distribution, e.g., the probability that a coin is fair, it may be worthwhile to carry through the normalization term and get actual posterior probabilities with "=".

[4]This rearrangement is a bit tricky. It involves recognizing the parameters of the posterior density function (in μ and h) as constructed from summary statistics of the data and parameters of the prior. At one point the rearrangement involves "completing the square."

The conditional posterior density of μ is

$$\pi(\mu|h,y) \propto \exp\left[-\frac{h_0 + hn}{2}\left(\mu - \frac{h_0\mu_0 + hn\bar{y}}{h_0 + hn}\right)^2\right],$$

which can be recognized as the density formula for a variable distributed $N\left[(h_0\mu_0 + hn\bar{y})/(h_0 + hn), (h_0 + hn)^{-1}\right]$. That is, the posterior distribution of μ is distributed normally with $E(\mu) = (h_0\mu_0 + hn\bar{y})/(h_0 + hn)$ and $\sigma_\mu^2 = (h_0 + hn)^{-1}$.

A reason to choose a normal distribution for the prior of μ and a Gamma distribution for the prior of h is that, for the parameters of a normal distribution (which define the likelihood function), these priors generate a normal distribution for the posterior of μ and a Gamma distribution for the posterior of h. When the posterior distribution uses the same functional form (with new parameter values, of course) as the prior, they are said to be "conjugate" and the problem is more tractable.

In this example it is reasonably easy to trace how the data, initially embedded only in the likelihood function, and the prior combine to form the posterior. We can examine the expected value of the mean μ in more detail:

$$E(\mu) = (h_0\mu_0 + hn\bar{y})/(h_0 + hn)$$

and rewrite this term to highlight how the posterior $E(\mu)$ is a weighted average[5] of the prior estimate of the mean, μ_0, and the empirical (maximum likelihood) estimate of the mean, $\bar{y}$:

$$E(\mu) = \frac{h_0}{h_0 + hn}\mu_0 + \frac{hn}{h_0 + hn}\bar{y}.$$

Note in particular how the relative weight on the second term depends on the sample size n. As the sample size gets larger and larger, the second weight approaches 1 (and the first weight approaches 0); hence the empirical estimate of the mean increasingly dominates the prior estimate as the sample gets large.[6]

Note also that the estimate of the posterior spread of beliefs about μ is $\sigma_\mu^2 = (h_0 + hn)^{-1} = \frac{1}{1/\sigma_{y0}^2 + n/\sigma_y^2}$, which looks something like the frequentist CLT estimate $\sigma_\mu^2 = \sigma_y^2/n$ (modified slightly by the inclusion of the prior estimate $h_0 = 1/\sigma_{y0}^2$).

In some sense, we are now done because we have written down a concise analytic expression for the posterior. But suppose that the posterior was very messy, and we were unable to write down an analytic expression. For example,

---

[5]Recall that the weights in a weighted average sum to 1 (in this case $\frac{h_0}{h_0 + hn} + \frac{hn}{h_0 + hn} = 1$) and express how much importance each of the terms being averaged carries in the average.

[6]I am less familiar with the Gamma distribution but the terms can be similarly interpreted to see the growing importance of the empirical estimate of the variance relative to the prior distribution.

the two parameters might not be independently distributed or the combination of the prior with the likelihood might be messy. In this case, we can use simulation to generate a posterior distribution for the parameters.[7]

Below is a Gibbs algorithm that randomly draws a new value of each parameter in successive steps. Note the iteration between the previous value of one parameter and the other. Some key elements of the simulation approach are the use of one previous value and the iterated conditionality, i.e., the next draw of one parameter depends on and only on one previous draw of the other parameter.

1. Choose a starting value for $\mu = \mu^{(0)}$.[8]

2. Sample $h^{(1)}$ from $G(\alpha_1/2, \delta_1/2)$, where $\alpha_1 = \alpha_0 + n$ and $\delta_1 = \delta_0 + \sum(y_i - \mu^{(0)})^2$.

3. At the gth iteration, draw
   $\mu^{(g)}$ from $N\left[(h_0\mu_0 + h^{(g-1)}n\bar{y})/(h_0 + h^{(g-1)}n), (h_0 + h^{(g-1)}n)^{-1}\right]$
   $h^{(g)}$ from $G\left[\alpha_1/2, \left(\delta_0 + \sum(y_i - \mu^{(g)})^2\right)/2\right]$

After discarding a number of burn-in draws, store the $\mu^{(g)}$ and $h^{(g)}$. These constitute the simulated distributions of μ and h and can be analyzed for means, credibility intervals (analogous to confidence intervals in frequentist statistics), etc. ℝ code and plots of the prior and posterior distributions follow.

```
1   ## Run with  R CMD BATCH --vanilla gibbs-normal-parameters.R &
2
3   ## Birth weight data from UC Berkeley
4   babiesI <- read.table("http://www.stat.berkeley.edu/users/statlabs/data/babiesI.data",header=TRUE)
5   ## babiesI <- read.table("babiesI.data",header=TRUE)
6
7   ## 2000 burn-in and 5000 post-burn-in iterations
8   burnin <- 2000
9   h <- rep(NA, 7000)
10  mu <- rep(NA, 7000)
11  delta <- rep(NA, 7000)
12
13  ## The truth
14  with(babiesI,mean(bwt))
15  with(babiesI,1/var(bwt))
16  with(babiesI,sd(bwt))
17
18  ## Draw a sample from the population
19  bwt_smpl <- with(babiesI, sample(bwt, size=20, replace = FALSE))
20
21  (n <- length(bwt_smpl))
22  (ybar <- mean(bwt_smpl))
23  sd(bwt_smpl)
24
25  ## Priors
26  mu0 <- 110
27  alpha0 <- 30
28  delta0 <- 6000
29  (h0 <- (alpha0/2) / (delta0/2))
```

---

[7]Note that we got the parameters by analytic means, not estimation by simulation, which is a different method. It's the distribution of the posterior which we will ostentatiously get by simulation using the analytically derived parameters.

[8]If desired the sampling can begin with $h^{(0)}$ and the algorithm modified accordingly.

```
30  ( 1 / h0 )
31  sqrt ( 1 / h0 )
32
33  alpha1 <- alpha0 + n
34  delta[1] <- delta0 + sum((bwt_smpl - mu0)^2)
35  h[1] <- rgamma(n=1,shape=alpha1/2,rate=delta[1]/2)
36  mu[1] <- rnorm(n=1,mean = (h0*mu0 + h0*n*ybar) / (h0 + h0*n), sd=sqrt ( 1 / (h0 + h0*n ) ) )
37
38  ## Here are the draws from the simulated distribution
39  for(i in 2:length(mu)) {
40    mu[i] <- rnorm(n=1,
41                   mean = (h0*mu0 + h[i-1]*n*ybar) / (h0 + h[i-1]*n),
42                   sd=sqrt ( 1 / (h0 + h[i-1] * n ) ) )
43    delta[i] = (delta0 + sum((bwt_smpl - mu[i])^2))
44    h[i] <- rgamma(n=1,shape=alpha1/2,rate= delta[i] / 2)
45  }
46
47  ## Report the mean and the 95 percent credibility interval of the
48  ## posterior of mu
49  mean(mu[(burnin+1):length(mu)])
50  quantile(mu[(burnin+1):length(mu)], probs=c(0.025,0.975))
51
52  ## Report the mean and 95 percent c.i. of the posterior of the
53  ## precision, the variance, and the standard deviation
54  mean(h[(burnin+1):length(h)])
55  quantile(h[(burnin+1):length(h)], probs=c(0.025,0.975))
56
57  (1/mean(h[(burnin+1):length(h)]))
58  quantile((1/h[(burnin+1):length(h)]), probs=c(0.025,0.975))
59
60  sqrt(1/mean(h[(burnin+1):length(h)]))
61  quantile(sqrt(1/h[(burnin+1):length(h)]), probs=c(0.025,0.975))
62
63  ## Prior of mean (mu)
64  pi.mu0 <- dnorm(x.mu <- seq(from=70,to=150,by=(150-70) / (length(mu)-burnin) )
    , mean=mu0, sd=sqrt(1/h0))
65  ## Posterior of mean (mu) with prior overlay
66  hist(mu[(burnin+1):length(mu)],xlim=c(70,150),main="",xlab="")
67  lines(x.mu,30000*pi.mu0)
68  title(expression(paste("Prior␣and␣Posterior␣Distribution␣of␣", mu)),xlab=expression(mu))
69
70  ## Prior of precision (h)
71  pi.h0 <- dgamma(x.h <- seq(from=(1/30)^2,to=(1/10)^2,by=((1/10)^2-(1/30)^2)/ (length(mu)-burnin) ),
72                  shape=alpha0 / 2, rate=delta0 / 2)
73  ## Posterior of precision (h) with prior overlay
74  hist(h[(burnin+1):length(h)],xlim=c((1/30)^2,(1/10)^2),main="",xlab="")
75  lines(x.h,2.7*pi.h0)
76  title(expression(paste("Prior␣and␣Posterior␣Distribution␣of␣h")),xlab=expression(paste("h")))
77
78  ## Prior of standard deviation (sqrt(1/h))
79  ## Posterior of standard deviation (sqrt(1/h)) with prior overlay
80  hist(sqrt(1/h[(burnin+1):length(h)]),xlim=c(10,30),main="",xlab="")
81  lines(sqrt(1/x.h),2.5*pi.h0)
82  title(expression(paste("Prior␣and␣Posterior␣Distribution␣of␣", sigma)),xlab=expression(sigma))
```