Notes for 751-752
Sections 23

Martin Lindquist*

December 11, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

23 Bayes analysis

23.1 Introduction to Bayesian analysis

Bayesian analysis is a form of statistical inference that relies on Baye's rule. The general version of Baye's rule states that

$$f(y|x) = f(x|y)f(y)/f(x)$$

where we're using f (loosely) as the appropriate density, mass function or probability and x and y represent random variables or events.

In the context of Bayesian analysis, Baye's rule is used in the following way. Let $\mathcal{L}(\theta; y)$ be the likelihood associated with data, y, and parameter θ. We codify our prior knowledge about θ with a prior distribution, $\pi(\theta)$. Then, a Bayesian analysis is performed via the posterior distribution

$$\pi(\theta \mid y) = f(y \mid \theta)\pi(\theta)/f(y) \propto_\theta f(y \mid \theta)\pi(\theta) \propto_\theta \mathcal{L}(\theta; y)\pi(\theta).$$

Therefore, one obtains the posterior, up to multiplicative constants, by multiplying the likelihood times the prior.

Coupled with Bayesian analysis is Bayesian interpretation of the probabilities. The prior is viewed as a belief and the posterior is then an updated belief coupling the objective evidence (the likelihood) with the subjective belief (the prior). By viewing probabilities as personal quantifications of beliefs, a Bayesian can talk about the probability of things that frequentists cannot. So, for example, if I roll a die and don't show you the result, as a Bayesian you can say that the probability that this specific roll is a six is one sixth. As a frequentist, in contrast, must say that in one sixth of repetitions of this experiment, the result will be a six. To a frequentist, this specific roll is either six or not.

This distinction in probabilistic interpretation has consequences in statistical interpretations. For example in diagnostic tests, a Bayesian can talk about the probability that a person has a disease, whereas a frequency interpretation relies on the percentage of diseased people in a population.

Personally, I've never minded either interpretation, but to many, the Bayesian interpretation seems more natural. In contrast, many practitioners dislike Bayesian analysis because of the prior specification, and the heavy reliance on fully specified models.

It should be noted that the discussion up to this point contrasted classical frequency thinking with classical subjective Bayesian thinking. In fact, most modern applied statisticians use hybrid approaches. They might, for example, develop a procedure with Bayesian tools (the manipulation of conditional distributions with priors on the parameters), but evaluate the procedure using frequency error rates. For all intents and purposes, such a procedure is frequentist, just developed with a Bayesian mindset. In contrast, many frequency statistical practitioners interpret their results with an approximate Bayesian mindset. Such procedures are simply Bayesian without the formalism. Even between these approaches there's continuous shades of gray. Therefore, saying a modern statistician is either Bayesian or frequentist is usually misleading, unless that

person does research or writes on statistical foundations. Nonetheless, foundational thinking is useful for understanding and clarifying thinking. It's worth reading and internalizing the literature on foundations for this reason alone. It's a lot like working on core drills to get better at a sport. That and it's quite a bit of fun!

Finally, it should also be noted that Bayes versus frequentist is far from the only schism in statistical foundations. Personally, I find the distinction between direct use of the design in frequency analysis to obtain robustness, like is often done is randomization testing and survey sampling, versus fully specified modeling a larger distinction than how one uses the model (Bayes versus frequentist). In addition, causal analysis versus association (non-causal) analysis forms a large distinction and one can perform Bayes or frequentists causal analysis and non-causal analyses. Furthermore, the likelihood paradigm offers a third inferential technique given a model over Bayes and frequency interpretations.

23.2 Basic Bayesian models

23.2.1 The Binomial model

We'll begin our discussion of Bayesian models by using some count outcome cases to build intuition. First, consider a series of coin flips, $X_1, \ldots, X_n \sim$ Bernoulli(θ). The likelihood associated with this experiment is

$$\mathcal{L}(\theta) \propto \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} = \theta^x(1-\theta)^{n-x}$$

where $x = \sum_i x_i$. Notice the likelihood depends only on the total number of successes. Consider putting a Beta(α, β) prior on θ. This can be written as follows:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $0 \leq \theta \leq 1$, and shape parameters α, $\beta > 0$. Here the terms α and β are refered to as hyperparameters.

Then, the posterior distribution is given by

$$\pi(\theta \mid x) \propto_\theta \mathcal{L}(\theta) \times \pi(\theta) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

therefore the posterior distribution is Beta$(x+\alpha, n-x+\beta)$. The posterior mean is

$$E[\theta \mid x] = \frac{x+\alpha}{n+\alpha+\beta} = \delta\hat{p} + (1-\delta)\frac{\alpha}{\alpha+\beta}$$

Therefore, the posterior mean is a weighted average of the MLE ($\hat{p}$) and the prior mean $\frac{\alpha}{\alpha+\beta}$. The weight is

$$\delta = \frac{n}{n+\alpha+\beta}.$$

Notice that, as $n \to \infty$ for fixed α and β, $\delta \to 1$ and the MLE dominates. That is, as we collect more data, the prior becomes less relevant and the data, in the form of the likelihood, dominates. On the other hand,

for fixed n, as either α or β go to infinity (or both), the prior dominates ($\delta \to 0$). For the Beta distribution α or β going to infinity makes the distribution much more peaked. Thus, if we are more certain of our prior distribution, the data matters less.

In Bayesian analysis, if the posterior is in the same family as the distribution, the prior is called a conjugate prior for the likelihood function. The Beta distribution is a conjugate prior for the Binomial distribution, as it gives rise to a posterior that follows a Beta distribution.

23.2.2 The Poisson model

Let $X \sim \text{Poisson}(t\lambda)$. Then

$$\mathcal{L}(\lambda) \propto \lambda^x e^{-t\lambda}.$$

Consider putting a Gamma(α, τ^{-1}) prior on λ. This can be written as follows:

$$p(\theta) = \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}$$

for $\theta > 0$, and shape parameters α, $\beta > 0$.

The posterior is the given by

$$\pi(\lambda \mid x) \propto \lambda^{x+\alpha-1} e^{-\lambda(t+\tau)}$$

and thus the posterior is Gamma$(x + \alpha, (t + \tau)^{-1})$. Because of the inversion of the second scale parameter of the Gamma, often Bayesians specify it in the terms of the inverse (as in Gamma$(x + \alpha, t + \tau)$). Often to avoid confusion, the mean of the gamma will be given to ensure no confusion over the parameterization.

The posterior mean is:

$$E[\lambda \mid x] = \frac{x+\alpha}{t+\tau} = \delta\hat{\lambda} + (1-\delta)\frac{\alpha}{\tau}$$

where $\hat{\lambda} = x/t$ is the MLE (the observed rate) and α/τ is the prior estimate. In this case

$$\delta = \frac{t}{t+\tau}$$

so that as $t \to \infty$ the MLE dominates while the prior dominates as $\tau \to \infty$.

23.2.3 The Normal model

Let us now turn our focus to the Normal model and discuss how to perform posterior inference on the population mean and variance. We begin with the case where the mean μ is the parameter of interest, and the variance σ^2 is known. Consider the following model:

$$\begin{aligned} y_1 \dots y_n \mid \mu &\sim N(\mu, \sigma^2) \\ \mu &\sim N(\mu_0, \tau_0^2) \end{aligned}$$

We know seek to compute $p(\mu|\mathbf{y})$.

Retaining only terms involving μ we have that the log of $f(\mu \mid \mathbf{y})$ is given by:

$$\begin{aligned}
\log(f(\mathbf{y} \mid \mu)) \quad &+ \quad \log(f(\mu)) \\
&\propto \quad -\frac{1}{2\sigma^2}\sum(y_i-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2 \\
&\propto \quad \mu n\bar{y}/\sigma^2 - \mu^2 n/(2\sigma^2) - \mu^2/(2\tau_0^2) + \mu\mu_0/\tau_0^2 \\
&= \quad \mu\left(\frac{\bar{y}}{\sigma^2/n} + \frac{\mu_0}{\tau_0^2}\right) - \frac{\mu^2}{2}\left(\frac{1}{\sigma^2/n} + \frac{1}{\tau_0^2}\right)
\end{aligned}$$

This can be recognized as the log density of a normally distributed random variable with variance

$$Var(\mu \mid \mathbf{y}) = \left(\frac{1}{\sigma^2/n} + \frac{1}{\tau_0^2}\right)^{-1} = \frac{\tau_0^2\sigma^2/n}{\sigma^2/n + \tau_0^2}$$

and mean

$$E[\mu \mid \mathbf{y}] = \left(\frac{1}{\sigma^2/n} + \frac{1}{\tau_0^2}\right)^{-1}\left(\frac{\bar{y}}{\sigma^2/n} + \frac{\mu_0}{\tau_0^2}\right).$$

Note, we can express the expected value as follows:

$$E[\mu \mid \mathbf{y}] = p\bar{y} + (1-p)\mu_0$$

where

$$p = \frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}.$$

Thus $E[\mu \mid \mathbf{y}]$ is a mixture of the empirical mean and the prior mean. How much the means are weighted depends on the ratio of the variance of the mean (σ^2/n) and the prior variance (τ_0^2). As we collect more data ($n \to \infty$), or if the data is not noisy ($\sigma \to 0$) or we have a lot of prior uncertainty ($\tau_0 \to \infty$) the empirical mean dominates. In contrast as we are more certain a priori ($\tau_0 \to 0$) the prior mean dominates.

Suppose that $\theta = (\theta_1, \ldots \theta_p)$, i.e., we are interested in working with multiple parameters simultaneously. The posterior density is now given by

$$f(\theta_1, \ldots \theta_p|y_1, \ldots y_n) \propto f(y_1, \ldots y_n|\theta)f(\theta)$$

To perform inference on a single parameter we need to find the marginal posterior for the parameter of interest. Theoretically, we can compute the marginal posterior as follows:

$$f(\theta_1|y_1, \ldots y_n) = \int \cdots \int f(\theta_1, \ldots \theta_p|y_1, \ldots y_n)d\theta_2 \ldots d\theta_p$$

In practice, it may not be feasible to compute this integral. In these cases using Monte Carlo methods is an option.

Revisiting the Normal model, in most situations we do not know σ^2 any more than we know μ. Thus, we thus need to expand our model to allow for joint inference for the mean and variance. Note, in the two parameter setting, the form of the joint posterior for μ and σ^2 would look as follows:

$$f(\mu, \sigma^2|y_1, \ldots y_n) \propto f(y_1, \ldots y_n|\mu, \sigma^2)\ f(\mu, \sigma^2).$$

We have now explicitly noted that σ^2 is also an unknown quantity, by including it in the prior distribution. Therefore, we now need to specify a joint prior for both μ and σ^2, and not just a prior for μ.

Let us begin by defining $\tau = 1/\sigma^2$, which we will work with instead of σ^2 to simplify calculations. We can now write the joint prior as follows:

$$f(\mu, \tau) = f(\mu|\tau)f(\tau)$$

A reasonable prior for μ given τ is from the Normal family. Similarly, one can show that the gamma family is a conjugate class of densities for the precision. This suggests the following model:

$$\begin{aligned} y_1, \ldots y_n|\mu, \tau &\sim N(\mu, 1/\tau) \\ \mu|\tau &\sim N(\mu, 1/(\kappa_0\tau)) \\ \tau &\sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

In this example, the joint prior follows a so-called normal-gamma distribution, i.e.

$$\begin{aligned} f(\mu, \tau) &\sim NG(\mu_0, \kappa_0, \nu_0, \sigma_0^2) \\ &\equiv N(\mu|\mu_0, 1/(\kappa_0\tau)) \times \text{gamma}(\tau|\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

There are four hyperparameters: μ_0, κ_0, ν_0, and σ_0^2. It turns out this is a conjugate prior for the two parameter Normal model. Thus, we can show that the joint posterior can be expressed as follows:

$$f(\mu, \tau|y_1, \ldots y_n) \sim NG(\mu, \tau|\mu_n, \kappa_n, \nu_n, \sigma_n^2)$$

where

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2 \end{aligned}$$

Note, here the term μ_n corresponds to a weighted average as before. The term $\nu_n\sigma_n^2$ is the sum of the sample sum of squares, the prior sum of squares, plus an addititional uncertainty term that arises due to the difference between the sample and prior mean.

We would like to make inferences about the marginal distributions $f(\mu|y)$ and $f(\sigma^2|y)$ rather than the conditional distribution $f(\mu, \sigma^2|y)$. The marginal posterior for τ is given by

$$\tau \mid y_1, \ldots y_n \sim \text{Gamma}(\nu_n/2, \nu_n\sigma_n^2/2).$$

The marginal posterior for μ is given by

$$\mu \mid y_1, \ldots y_n \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n).$$

which is a kernel of a tdistribution with ν_n degrees of freedom, centered at μ_n and with scale parameter σ_n^2/κ_n.

23.3 Bayesian Linear Models

Recall our standard Gaussian linear model, where we can write:

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

. Consider three common prior specifications:

1. $\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Sigma}_0)$ and $\sigma^{-2} \sim \text{Gamma}(\alpha_0, \tau_0^{-1})$.
2. $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ and $\sigma^{-2} \sim \text{Gamma}(\alpha_0, \tau_0^{-1})$.
3. $(\boldsymbol{\beta}, \sigma^2) \sim \sigma^{-2}$.

The distinction between the first case and the second is the inclusion of σ^2 in the prior specification for $\boldsymbol{\beta}$. This is useful for making all posterior distributions tractable, including that of $\boldsymbol{\beta}$ integrated over σ^2. However, it may or may not reflect the desired prior distribution.

The second specification has tractable full conditionals. That is, we can easily figure out $\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}$ and $\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}$. However, the posterior marginals of the parameters ($\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}$ in particular) are not tractable. This posterior is often explored using Monte Carlo.

The third specification is also completely tractable, even thoughhe final prior specification is not a proper density. It doesn't have a defined integral for the elements of $\boldsymbol{\beta}$ from $-\infty$ to ∞ and for $0 \leq \sigma^2 < \infty$. However, proceeding as if it were a proper density yields a proper distribution for the posterior. Such "improper" priors are often used to specify putatively uninformative distributions that yield valid posteriors. In this case, the posterior has the property of the posterior mode being centered around $\hat{\boldsymbol{\beta}}$.

23.4 Monte Carlo sampling

Even though many of our Bayesian models are completely tractable, we will explore the posteriors via Monte Carlo. The reason for this is to get students familiar with Monte Carlo so that they can apply it in the more complex settings that they are likely to encounter in practice. Specifically, usually fully tractable posteriors are more of an exception than the rule. For the most part, for linear models, one should use the fully tractable results as they are much faster.

We now give some strategies for Monte Carlo sampling from a posterior.

23.4.1 Sequential Monte Carlo

Notice for three variables, X, Y and Z, sampling $f_z(z)$, $f_{y|z}(y)$ and $f_{x|y,z}(x \mid y, z)$ yields a multivariate draw from the joint distribution $f(x, y, z)$. So, for example, consider setting 3 of our prior specifications

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X} \sim N(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}\sigma^2) \text{ and } \sigma^{-2} \mid \mathbf{y}, \mathbf{X} \sim \text{Gamma}\{(n-p)/2, 2/(n-p)S^2\}$$

Notice that $E[\sigma^{-2} \mid \mathbf{X}, \mathbf{y}] = 1/S^2$. To simulate from this distribution, we first simulate from $\sigma^{-2} \mid \mathbf{X}, \mathbf{y}$ then plug that simulation into $\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}$ and simulate an $\boldsymbol{\beta}$. The pair is a draw from the joint posterior distribution of $\boldsymbol{\beta}$ and σ^{-2}.

23.5 Gibbs sampling

Consider again our three random variables. Suppose that an initial value of x and y, say $x^{(1)}$ and $y^{(1)}$ was obtained. Then consider simulating

1. $z^{(1)} \sim f_{z|x,y}(z \mid x^{(1)}, y^{(1)})$
2. $x^{(2)} \sim f_{x|y,z}(x \mid y^{(1)}, z^{(2)})$
3. $y^{(2)} \sim f_{y|x,z}(y \mid x^{(2)}, z^{(2)})$
4. $z^{(2)} \sim f_{z|x,y}(z \mid x^{(2)}, y^{(2)})$
5. $x^{(3)} \sim f_{x|y,z}(x \mid y^{(2)}, z^{(3)})$

and so on. In other words, always update a simulated variable using the most recently simulated version of the other variables. In fact, one need not use the full conditionals. Any collection of conditionals would work. Moreover, any random order works, or even randomizing the order each iteration. However, some conditions have to be met for the asymptotics of the sampler to work. For example, you have to update every variable infinitely often and the whole space has to be explorable by the sampler. If the conditions are

met, the sampler is a Markov chain whose stationary distribution, i.e. the limiting distribution, is $f(x, y, z)$. Moreover, there's lots of results saying that you can use the ouput of the sampler in much the same way one uses iid samples. For example approximating posterior means with the average of the simulated variables. However, the Markovian nature of the sampler makes using the samples a little trickier. One could try to combat this by running the chain for a while and throwing out all of the early simulations used to "burn in" the sampler. This throws away a lot of data. Our preferred method is to use good starting values (why not start at the MLE?) and use all of the simulated data.

Let's illustrate the sampler with prior specification 2. Consider the simplified model:

$$\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X}, \theta \sim N(\mathbf{X}\boldsymbol{\beta}, \theta^{-1}\mathbf{I}) \text{ and } \boldsymbol{\beta} \sim N(\mathbf{0}, \psi^{-1}\mathbf{I}) \text{ and } \theta \sim \text{Gamma}(\alpha/2, \tau^{-1}).$$

Under this specification, the full conditionals are:

$$\begin{aligned}\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \theta &\sim N\{(\mathbf{X}'\mathbf{X}\theta + \psi\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, (\mathbf{X}\mathbf{X}\theta + \psi\mathbf{I})^{-1}\} \\ \theta \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\beta} &\sim \text{Gamma}\{(n+\alpha)/2, 2(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau)^{-1}\}\end{aligned}$$

23.5.1 Coding Example

```
data("mtcars")
y = mtcars$mpg - mean(mtcars$mpg)
x = cbind(1, mtcars$wt - mean(mtcars$wt))

n = length(y)
p = ncol(x)

fitML  = lm(y ~ x - 1)


xtx = t(x) %*% x
xty = as.vector(t(x) %*% y)

nosim = 10000

rmvnorm = function(mu, Sigma) as.vector(mu + chol(Sigma) %*% rnorm(length(mu)))

thetaCurrent = 1 / summary(fitML)$sigma^2
beta = NULL
theta = thetaCurrent * 100
```

```
psi = .01
alpha = .01
tau = .01 * summary(fitML)$sigma^2
for (i in 1 : nosim){
  V = solve(xtx * thetaCurrent + psi * diag(1, p, p))
  mu = V %*% xty * thetaCurrent
  betaCurrent = rmvnorm(mu, V)
  sumesq = sum((y - x %*% betaCurrent)^2)
  thetaCurrent = rgamma(1, (n + alpha) / 2, rate = (sumesq + tau)/2)
  theta = c(theta, thetaCurrent)
  beta = rbind(beta, betaCurrent)
}
sigma = sqrt(1/ theta)
quantile(beta[,1], c(.025, .975))
quantile(beta2[,2], c(.025, .975))
quantile(sigma, c(0.025, .975))
```