Chapter 12

Bayesian Inference

This chapter covers the following topics:

- Concepts and methods of Bayesian inference.
- Bayesian hypothesis testing and model comparison.
- Derivation of the Bayesian information criterion (BIC).
- Simulation methods and Markov chain Monte Carlo (MCMC).
- Bayesian computation via variational inference.
- Some subtle issues related to Bayesian inference.

12.1 What is Bayesian Inference?

There are two main approaches to statistical machine learning: *frequentist* (or classical) methods and *Bayesian* methods. Most of the methods we have discussed so far are frequentist. It is important to understand both approaches. At the risk of oversimplifying, the difference is this:

Frequentist versus Bayesian Methods

- In frequentist inference, probabilities are interpreted as long run frequencies. The goal is to create procedures with long run frequency guarantees.
- In Bayesian inference, probabilities are interpreted as subjective degrees of belief. The goal is to state and analyze your beliefs.

Some differences between the frequentist and Bayesian approaches are as follows:

	Frequentist	Bayesian
Probability is:	limiting relative frequency	degree of belief
Parameter θ is a:	fixed constant	random variable
Probability statements are about:	procedures	parameters
Frequency guarantees?	yes	no

To illustrate the difference, consider the following example. Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$. We want to provide some sort of interval estimate C for θ.

Frequentist Approach. Construct the confidence interval

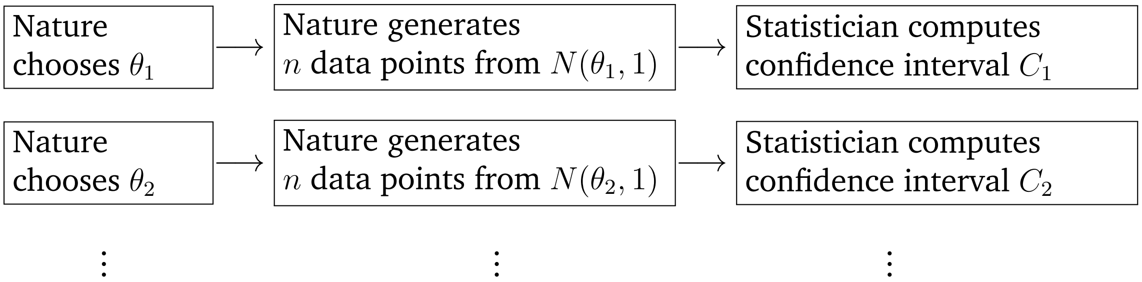
$$C = \left[\overline{X}_n - \frac{1.96}{\sqrt{n}},\ \overline{X}_n + \frac{1.96}{\sqrt{n}}\right].$$

Then

$$\mathbb{P}_\theta(\theta \in C) = 0.95 \quad \text{for all } \theta \in \mathbb{R}.$$

The probability statement is about the random interval C. The interval is random because it is a function of the data. The parameter θ is a fixed, unknown quantity. The statement means that C will trap the true value with probability 0.95.

To make the meaning clearer, suppose we repeat this experiment many times. In fact, we can even allow θ to change every time we do the experiment. The experiment looks like this:


Nature chooses θ_1
Nature generates n data points from $N(\theta_1, 1)$
Statistician computes confidence interval C_1
Nature chooses θ_2
Nature generates n data points from $N(\theta_2, 1)$
Statistician computes confidence interval C_2
⋮
⋮
⋮

We will find that the interval C_j traps the parameter θ_j, 95 percent of the time. More precisely,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\theta_i \in C_i) \geq 0.95 \tag{12.1}$$

almost surely, for any sequence $\theta_1, \theta_2, \ldots$.

Bayesian Approach. The Bayesian treats probability as beliefs, not frequencies. The unknown parameter θ is given a prior distributon $\pi(\theta)$ representing his subjective beliefs about θ. After seeing the data $X_1, \ldots, X_n$, he computes the posterior distribution for θ given the data using Bayes theorem:

$$\pi(\theta|X_1, \ldots, X_n) \propto \mathcal{L}(\theta)\pi(\theta) \tag{12.2}$$

where $\mathcal{L}(\theta)$ is the likelihood function. Next we finds an interval C such that

$$\int_C \pi(\theta|X_1, \ldots, X_n)d\theta = 0.95.$$

He can thn report that

$$\mathbb{P}(\theta \in C|X_1, \ldots, X_n) = 0.95.$$

This is a degree-of-belief probablity statement about θ given the data. It is not the same as (12.1). If we repeated this experient many times, the intervals would **not** trap the true value 95 percent of the time.

Frequentist inference is aimed at given procedures with frequency guarantees. Bayesian inference is about stating and manipulating subjective beliefs. In general, these are different, A lot of confusion would be avoided if we used $F(C)$ to denote frequency probablity and $B(C)$ to denote degree-of-belief probability. These are idfferent things and there is no reason to expect them to be the same. Unfortunately, it is traditional to use the same symbol, such as $\mathbb{P}$, to denote both types of probability which leads to confusion.

To summarize: Frequentist inference gives procedures with frequency probability guarantees. Bayesian inference is a method for stating and updating beliefs. A frequentist confidence interval C satisfies

$$\inf_\theta \mathbb{P}_\theta(\theta \in C) = 1 - \alpha$$

where the probability refers to random interval C. We call $\inf_\theta \mathbb{P}_\theta(\theta \in C)$ the coverage of the interval C. A Bayesian confidence interval C satisfies

$$\mathbb{P}(\theta \in C|X_1, \ldots, X_n) = 1 - \alpha$$

where the probability refers to θ. Later, we will give concrete examples where the coverage and the posterior probability are very different.

Remark. There are, in fact, many flavors of Bayesian inference. Subjective Bayesians interpret probability strictly as personal degrees of belief. Objective Bayesians try to find prior distributions that formally express ignorance with the hope that the resulting posterior is, in some sense, objective. Empirical Bayesians estimate the prior distribution from the data. Frequentist Bayesians are those who use Bayesian methods only when the resulting posterior has good frequency behavior. Thus, the distinction between Bayesian and frequentist inference can be somewhat murky. This has led to much confusion in statistics, machine learning and science.

12.2 Basic Concepts

Let $X_1, \ldots, X_n$ be n observations sampled from a probability density $p(\mathrm{x} \,|\, \theta)$. In this chapter, we write $p(\mathrm{x} \,|\, \theta)$ if we view θ as a random variable and $p(\mathrm{x} \,|\, \theta)$ represents the conditional probability density of X conditioned on θ. In contrast, we write $p_\theta(\mathrm{x})$ if we view θ as a deterministic value.

12.2.1 The Mechanics of Bayesian Inference

Bayesian inference is usually carried out in the following way.

Bayesian Procedure

1. We choose a probability density $\pi(\theta)$ — called the prior distribution — that expresses our beliefs about a parameter θ before we see any data.
2. We choose a statistical model $p(\mathrm{x} \,|\, \theta)$ that reflects our beliefs about x given θ.
3. After observing data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$, we update our beliefs and calculate the posterior distribution $p(\theta \,|\, \mathcal{D}_n)$.

By Bayes' theorem, the posterior distribution can be written as

$$p(\theta \,|\, X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n \,|\, \theta)\pi(\theta)}{p(X_1, \ldots, X_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta) \qquad (12.3)$$

where $\mathcal{L}_n(\theta) = \prod_{i=1}^n p(X_i \,|\, \theta)$ is the likelihood function and

$$c_n = p(X_1, \ldots, X_n) = \int p(X_1, \ldots, X_n \,|\, \theta)\pi(\theta)d\theta = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called the evidence.

We can get a Bayesian point estimate by summarizing the center of the posterior. Typically, we use the mean or mode of the posterior distribution. The posterior mean is

$$\overline{\theta}_n = \int \theta p(\theta \,|\, \mathcal{D}_n)d\theta = \frac{\displaystyle\int \theta \mathcal{L}_n(\theta)\pi(\theta)d\theta}{\displaystyle\int \mathcal{L}_n(\theta)\pi(\theta)d\theta}.$$

We can also obtain a Bayesian interval estimate. For example, for $\alpha \in (0,1)$, we could find a and b such that

$$\int_{-\infty}^{a} p(\theta \,|\, \mathcal{D}_n)\, d\theta = \int_{b}^{\infty} p(\theta \,|\, \mathcal{D}_n)\, d\theta = \alpha/2.$$

Let $C = (a,b)$. Then

$$\mathbb{P}(\theta \in C \,|\, \mathcal{D}_n) = \int_a^b p(\theta \,|\, \mathcal{D}_n)\, d\theta = 1 - \alpha,$$

so C is a $1-\alpha$ Bayesian posterior interval or credible interval. If θ has more than one dimension, the extension is straightforward and we obtain a credible region.

Example 205. Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ where $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(\theta)$. Suppose we take the uniform distribution $\pi(\theta) = 1$ as a prior. By Bayes' theorem, the posterior is

$$p(\theta \,|\, \mathcal{D}_n) \propto \pi(\theta)\mathcal{L}_n(\theta) = \theta^{S_n}(1-\theta)^{n-S_n} = \theta^{S_n+1-1}(1-\theta)^{n-S_n+1-1}$$

where $S_n = \sum_{i=1}^n X_i$ is the number of successes. Recall that a random variable θ on the interval $(0,1)$ has a Beta distribution with parameters α and β if its density is

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

We see that the posterior distribution for θ is a Beta distribution with parameters $S_n + 1$ and $n - S_n + 1$. That is,

$$p(\theta \,|\, \mathcal{D}_n) = \frac{\Gamma(n+2)}{\Gamma(S_n+1)\Gamma(n-S_n+1)}\theta^{(S_n+1)-1}(1-\theta)^{(n-S_n+1)-1}.$$

We write this as

$$\theta \,|\, \mathcal{D}_n \sim \mathrm{Beta}(S_n+1, n-S_n+1).$$

Notice that we have figured out the normalizing constant without actually doing the integral $\int \mathcal{L}_n(\theta)\pi(\theta)\, d\theta$. Since a density function integrates to one, we see that

$$\int_0^1 \theta^{S_n}(1-\theta)^{n-S_n} = \frac{\Gamma(S_n+1)\Gamma(n-S_n+1)}{\Gamma(n+2)}.$$

The mean of a $\mathrm{Beta}(\alpha,\beta)$ distribution is $\alpha/(\alpha+\beta)$ so the Bayes posterior estimator is

$$\overline{\theta} = \frac{S_n+1}{n+2}.$$

It is instructive to rewrite $\overline{\theta}$ as

$$\overline{\theta} = \lambda_n \widehat{\theta} + (1-\lambda_n)\widetilde{\theta}$$

where $\widehat{\theta} = S_n/n$ is the maximum likelihood estimate, $\widetilde{\theta} = 1/2$ is the prior mean and $\lambda_n = n/(n+2) \approx 1$. A 95 percent posterior interval can be obtained by numerically finding a and b such that $\int_a^b p(\theta \,|\, \mathcal{D}_n)\, d\theta = .95$.

Suppose that instead of a uniform prior, we use the prior $\theta \sim \text{Beta}(\alpha, \beta)$. If you repeat the calculations above, you will see that $\theta \,|\, \mathcal{D}_n \sim \text{Beta}(\alpha + S_n, \beta + n - S_n)$. The flat prior is just the special case with $\alpha = \beta = 1$. The posterior mean in this more general case is

$$\overline{\theta} = \frac{\alpha + S_n}{\alpha + \beta + n} = \left(\frac{n}{\alpha + \beta + n}\right)\widehat{\theta} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\theta_0$$

where $\theta_0 = \alpha/(\alpha + \beta)$ is the prior mean.

An illustration of this example is shown in Figure 12.1. We use the Bernoulli model to generate $n = 15$ data with parameter $\theta = 0.4$. We observe $s = 7$. Therefore, the maximum likelihood estimate is $\widehat{\theta} = 7/15 = 0.47$, which is larger than the true parameter value 0.4. The left plot of Figure 12.1 adopts a prior $\text{Beta}(4, 6)$ which gives a posterior mode 0.43, while the right plot of Figure 12.1 adopts a prior $\text{Beta}(4, 2)$ which gives a posterior mode 0.67.

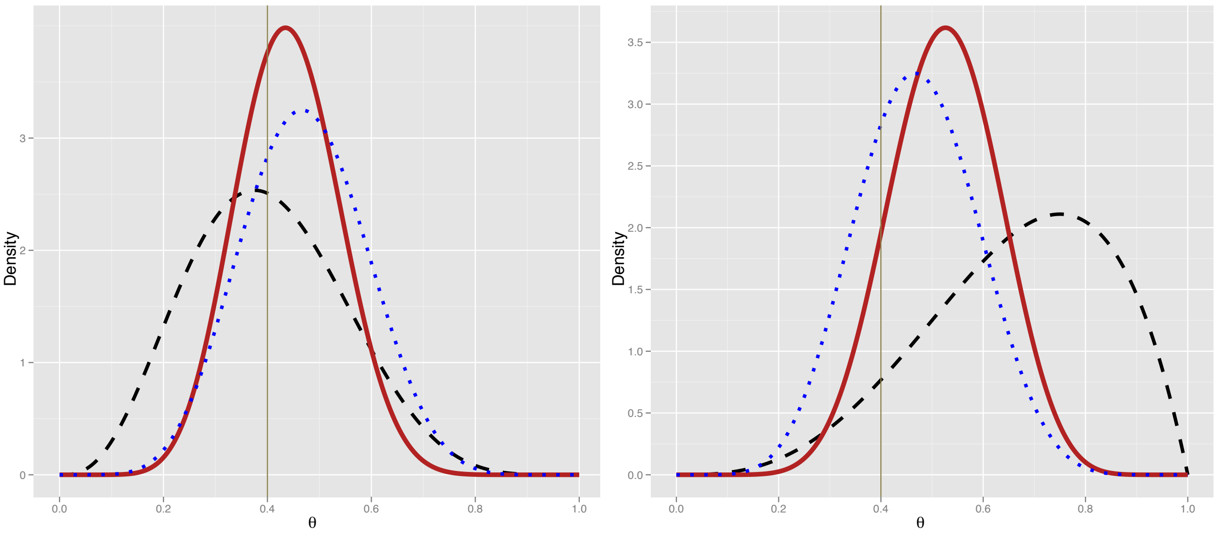


Figure 12.1: Illustration of Bayesian inference on Bernoulli data with two priors. The three curves are prior distribution (red-solid), likelihood function (blue-dashed), and the posterior distribution (black-dashed). The true parameter value $\theta = 0.4$ is indicated by the vertical line.

Example 206. Let $\boldsymbol{X} \sim \text{Multinomial}(n, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ be a K-dimensional parameter ($K > 1$). The multinomial model with a Dirichlet prior is a generalization of the Bernoulli model and Beta prior of the previous example. The Dirichlet distribution for

K outcomes is the exponential family distribution on the $K-1$ dimensional probability simplex[1] Δ_K given by

$$\pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \;=\; \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K \theta_j^{\alpha_j - 1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T \in \mathbb{R}_+^K$ is a non-negative vector of scaling coefficients, which are the parameters of the model. We can think of the sample space of the multinomial with K outcomes as the set of vertices of the K-dimensional hypercube $\mathbb{H}_K$, made up of vectors with exactly one 1 and the remaining elements 0:

$$\mathbf{x} = \underbrace{(0, 0, \ldots, 0, 1, 0, \ldots, 0)^T}_{K \text{ places}}.$$

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})^T \in \mathbb{H}_K$. If

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \quad \text{and} \quad \boldsymbol{X}_i \,|\, \boldsymbol{\theta} \sim \text{Multinomial}(\boldsymbol{\theta}) \;\text{ for } i = 1, 2, \ldots, n,$$

then the posterior satisfies

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \propto \mathcal{L}_n(\theta)\, \pi(\theta) \propto \prod_{i=1}^n \prod_{j=1}^K \theta_j^{X_{ij}} \prod_{j=1}^K \theta_j^{\alpha_j - 1} = \prod_{j=1}^K \theta_j^{\sum_{i=1}^n X_{ij} + \alpha_j - 1}.$$

We see that the posterior is also a Dirichlet distribution:

$$\boldsymbol{\theta} \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \;\sim\; \text{Dirichlet}(\alpha + n\overline{\boldsymbol{X}})$$

where $\overline{\boldsymbol{X}} = n^{-1} \sum_{i=1}^n \boldsymbol{X}_i \in \Delta_K$.

Since the mean of a Dirichlet distribution $\pi_\alpha(\boldsymbol{\theta})$ is given by

$$\mathbb{E}(\boldsymbol{\theta}) = \left(\frac{\alpha_1}{\sum_{i=1}^K \alpha_i}, \ldots, \frac{\alpha_K}{\sum_{i=1}^K \alpha_i} \right)^T,$$

the posterior mean of a multinomial with Dirichlet prior is

$$\mathbb{E}(\theta \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \left(\frac{\alpha_1 + \sum_{i=1}^n X_{i1}}{\sum_{i=1}^K \alpha_i + n}, \ldots, \frac{\alpha_K + \sum_{i=1}^n X_{iK}}{\sum_{i=1}^K \alpha_i + n} \right)^T.$$

[1] The probability simplex Δ_K is defined as

$$\Delta_K = \Big\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T \in \mathbb{R}^K \,|\, \theta_i \geq 0 \text{ for all } \; i \text{ and } \sum_{i=1}^K \theta_i = 1 \Big\}.$$

The posterior mean can be viewed as smoothing out the maximum likelihood estimate by allocating some additional probability mass to low frequency observations. The parameters $\alpha_1, \ldots, \alpha_K$ act as "virtual counts" that don't actually appear in the observed data.

An illustration of this example is shown in Figure 12.2. We use the multinomial model to generate $n = 20$ data points with parameter $\boldsymbol{\theta} = (0.2, 0.3, 0.5)^T$. We adopt a prior $\text{Dirichlet}(6, 6, 6)$. The contours of the prior, likelihood, and posterior with $n = 20$ observed data are shown in the first three plots in Figure 12.2. As a comparison, we also provide the contour of the posterior with $n = 200$ observed data in the last plot. From this experiment, we see that when the number of observed data is small, the posterior is affected by both the prior and the likelihood; when the number of observed data is large, the posterior is mainly dominated by the likelihood.

In the previous two examples, the prior was a Dirichlet distribution and the posterior was also a Dirichlet. When the prior and the posterior are in the same family, we say that the prior is conjugate with respect to the model; this will be discussed further below.

Example 207. Let $X \sim N(\theta, \sigma^2)$ and $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ be the observed data. For simplicity, let us assume that σ is known and we want to estimate $\theta \in \mathbb{R}$. Suppose we take as a prior $\theta \sim N(a, b^2)$. Let $\overline{X} = \sum_{i=1}^n X_i/n$ be the sample mean. In the Exercise, it is shown that the posterior for θ is

$$\theta \,|\, \mathcal{D}_n \sim N(\overline{\theta}, \tau^2) \tag{12.4}$$

where

$$\overline{\theta} = w\widehat{\theta} + (1 - w)a,$$

$$\widehat{\theta} = \overline{X}, \quad w = \frac{\frac{1}{\text{se}^2}}{\frac{1}{\text{se}^2} + \frac{1}{b^2}}, \quad \frac{1}{\tau^2} = \frac{1}{\text{se}^2} + \frac{1}{b^2},$$

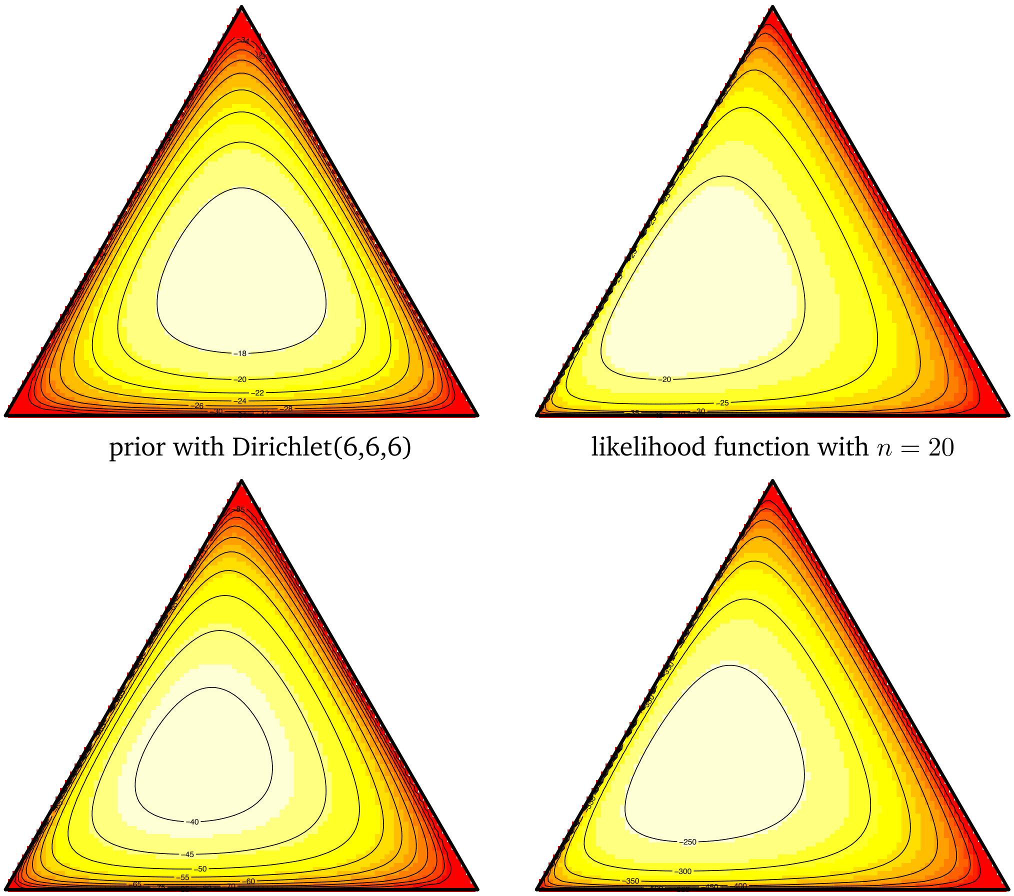
and $\text{se} = \sigma/\sqrt{n}$ is the standard error of the maximum likelihood estimate $\widehat{\theta}$. This is another example of a conjugate prior. Note that $w \to 1$ and $\tau/\text{se} \to 1$ as $n \to \infty$. So, for large n, the posterior is approximately $N(\widehat{\theta}, \text{se}^2)$. The same is true if n is fixed but $b \to \infty$, which corresponds to letting the prior become very flat.

Continuing with this example, let us find $C = (c, d)$ such that $\mathbb{P}(\theta \in C \,|\, \mathcal{D}_n) = 0.95$. We can do this by choosing c and d such that $\mathbb{P}(\theta < c \,|\, \mathcal{D}_n) = 0.025$ and $\mathbb{P}(\theta > d \,|\, \mathcal{D}_n) = 0.025$. More specifically, we want to find c such that

$$\mathbb{P}(\theta < c \,|\, \mathcal{D}_n) = \mathbb{P}\left(\frac{\theta - \overline{\theta}}{\tau} < \frac{c - \overline{\theta}}{\tau} \,\middle|\, \mathcal{D}_n \right) = \mathbb{P}\left(Z < \frac{c - \overline{\theta}}{\tau} \right) = 0.025$$

where $Z \sim N(0, 1)$ is a standard Gaussian random variable. We know that $\mathbb{P}(Z < -1.96) = 0.025$. So,

$$\frac{c - \overline{\theta}}{\tau} = -1.96$$



prior with Dirichlet(6,6,6)

likelihood function with $n = 20$

posterior distribution with $n = 20$

posterior distribution with $n = 200$

Figure 12.2: Illustration of Bayesian inference on multinomial data with the prior $\mathrm{Dirichlet}(6, 6, 6)$. The contours of the prior, likelihood, and posteriors are plotted on a two-dimensional probability simplex (Starting from the bottom left vertex of each triangle, clock-wisely the three vertices correspond to $\theta_1, \theta_2, \theta_3$). We see that when the number of observed data is small, the posterior is affected by both the prior and the likelihood; when the number of observed data is large, the posterior is mainly dominated by the likelihood.

implying that $c = \overline{\theta} - 1.96\tau$. By similar arguments, $d = \overline{\theta} + 1.96\tau$. So a 95 percent Bayesian credible interval is $\overline{\theta} \pm 1.96\,\tau$. Since $\overline{\theta} \approx \widehat{\theta}$ and $\tau \approx \mathrm{se}$ when n is large, the 95 percent Bayesian credible interval is approximated by $\widehat{\theta} \pm 1.96\,\mathrm{se}$ which is the frequentist confidence interval.

12.2.2 Bayesian Prediction

After the data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ have been observed, the Bayesian framework allows us to predict the distribution of a future data point X conditioned on $\mathcal{D}_n$. To do this, we first obtain the posterior $p(\theta \,|\, \mathcal{D}_n)$. Then

$$\begin{aligned} p(\mathrm{x} \,|\, \mathcal{D}_n) &= \int p(\mathrm{x}, \theta \,|\, \mathcal{D}_n) d\theta \\ &= \int p(\mathrm{x} \,|\, \theta, \mathcal{D}_n) p(\theta \,|\, \mathcal{D}_n) d\theta \\ &= \int p(\mathrm{x} \,|\, \theta) p(\theta \,|\, \mathcal{D}_n) d\theta. \end{aligned}$$

Where we use the fact that $p(\mathrm{x} \,|\, \theta, \mathcal{D}_n) = p(\mathrm{x} \,|\, \theta)$ since all the data are conditionally independent given θ. From the last line, the predictive distribution $p(\mathrm{x} \,|\, \mathcal{D}_n)$ can be viewed as a weighted average of the model $p(\mathrm{x} \,|\, \theta)$. The weights are determined by the posterior distribution of θ.

12.2.3 Inference about Functions of Parameters

Given the data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$, how do we make inferences about a function $\tau = g(\theta)$? The posterior CDF for τ is

$$H(t \,|\, \mathcal{D}_n) = \mathbb{P}(g(\theta) \leq t \,|\, \mathcal{D}_n) = \int_A p(\theta \,|\, \mathcal{D}_n) d\theta$$

where $A = \{\theta : \; g(\theta) \leq t\}$. The posterior density is $p(\tau \,|\, \mathcal{D}_n) = H'(\tau \,|\, \mathcal{D}_n)$.

Example 208. Under a Bernoulli model $X \sim \mathrm{Bernoulli}(\theta)$, let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ be the observed data and $\pi(\theta) = 1$ so that $\theta \,|\, \mathcal{D}_n \sim \mathrm{Beta}(S_n + 1, n - S_n + 1)$ with $S_n = \sum_{i=1}^n X_i$. We define $\psi = \log(\theta/(1-\theta))$. Then

$$\begin{aligned} H(t \,|\, \mathcal{D}_n) &= \mathbb{P}(\psi \leq t \,|\, \mathcal{D}_n) = \mathbb{P}\Big(\log\Big(\frac{\theta}{1-\theta}\Big) \leq t \,|\, \mathcal{D}_n\Big) \\ &= \mathbb{P}\Big(\theta \leq \frac{e^t}{1+e^t} \,|\, \mathcal{D}_n\Big) \\ &= \int_0^{e^t/(1+e^t)} p(\theta \,|\, \mathcal{D}_n) \, d\theta \\ &= \frac{\Gamma(n+2)}{\Gamma(S_n+1)\Gamma(n-S_n+1)} \int_0^{e^t/(1+e^t)} \theta^{S_n} (1-\theta)^{n-S_n} \, d\theta \end{aligned}$$

and

$$p(\psi \,|\, \mathcal{D}_n) = H'(\psi \,|\, \mathcal{D}_n) = \frac{\Gamma(n+2)}{\Gamma(S_n+1)\Gamma(n-S_n+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^{S_n} \left(\frac{1}{1+e^\psi}\right)^{n-S_n+2}$$

for $\psi \in \mathbb{R}$.

12.2.4 Multiparameter Problems

Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ be the observed data. Suppose that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ with some prior distribution $\pi(\boldsymbol{\theta})$. The posterior density is still given by

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}_n) \propto \mathcal{L}_n(\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

The question now arises of how to extract inferences about one single parameter. The key is to find the marginal posterior density for the parameter of interest. Suppose we want to make inferences about θ_1. The marginal posterior for θ_1 is

$$p(\theta_1 \,|\, \mathcal{D}_n) = \int \cdots \int p(\theta_1, \cdots, \theta_d \,|\, \mathcal{D}_n) d\theta_2 \ldots d\theta_d.$$

In practice, it might not be feasible to do this integral. Simulation can help: we draw randomly from the posterior:

$$\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^B \sim p(\boldsymbol{\theta} \,|\, \mathcal{D}_n)$$

where the superscripts index different draws. Each $\boldsymbol{\theta}^j$ is a vector $\boldsymbol{\theta}^j = (\theta_1^j, \ldots, \theta_d^j)^T$. Now collect together the first component of each draw: $\theta_1^1, \ldots, \theta_1^B$. These are a sample from $p(\theta_1 \,|\, \mathcal{D}_n)$ and we have avoided doing any integrals. One thing to note is, sampling B data from a multivariate distribution $p(\boldsymbol{\theta} \,|\, \mathcal{D}_n)$ is challenging especially when the dimensionality d is large. We will discuss this topic further in the section on Bayesian computation.

Example 209 (Comparing Two Binomials)**.** Suppose we have n_1 control patients and n_2 treatment patients and that X_1 is the number of survived patients in the control group; while x_2 is the number of survived patients in the treatment group. We assume the Binomial model:

$$X_1 \sim \text{Binomial}(n_1, \theta_1) \ \text{ and } \ X_2 \sim \text{Binomial}(n_2, \theta_2).$$

We want to estimate $\tau = g(\theta_1, \theta_2) = \theta_2 - \theta_1$.

If $\pi(\theta_1, \theta_2) = 1$, the posterior is

$$p(\theta_1, \theta_2 \,|\, X_1, X_2) \propto \theta_1^{X_1}(1-\theta_1)^{n_1 - X_1}\theta_2^{X_2}(1-\theta_2)^{n_2 - X_2}.$$

Notice that (θ_1, θ_2) live on a rectangle (a square, actually) and that

$$p(\theta_1, \theta_2 \,|\, X_1, X_2) = p(\theta_1 \,|\, X_1) p(\theta_2 \,|\, X_2)$$

where

$$p(\theta_1 \,|\, X_1) \propto \theta_1^{X_1}(1-\theta_1)^{n_1 - X_1} \ \text{ and } \ p(\theta_2 \,|\, X_2) \propto \theta_2^{X_2}(1-\theta_2)^{n_2 - X_2},$$

which implies that θ_1 and θ_2 are independent under the posterior. Also, $\theta_1 \,|\, X_1 \sim \text{Beta}(X_1 + 1, n_1 - X_1 + 1)$ and $\theta_2 \,|\, X_2 \sim \text{Beta}(X_2 + 1, n_2 - X_2 + 1)$. If we simulate $\theta_1^1, \ldots, \theta_1^B \sim \text{Beta}(X_1 + 1, n_1 - X_1 + 1)$ and $\theta_2^1, \ldots, \theta_2^B \sim \text{Beta}(X_2 + 1, n_2 - X_2 + 1)$, then $\tau_b = \theta_2^b - \theta_1^b$, $b = 1, \ldots, B$, is a sample from $p(\tau \,|\, X_1, X_2)$.

12.2.5 Flat Priors, Improper Priors, and "Noninformative" Priors

An important question in Bayesian inference is: where does one get the prior $\pi(\theta)$? One school of thought, called subjectivism says that the prior should reflect our subjective opinion about θ (before the data are collected). This may be possible in some cases but is impractical in complicated problems especially when there are many parameters. Moreover, injecting subjective opinion into the analysis is contrary to the goal of making scientific inference as objective as possible.

An alternative is to try to define some sort of "noninformative prior." An obvious candidate for a noninformative prior is to use a flat prior $\pi(\theta) \propto \mathrm{constant}$. In the Bernoulli example, taking $\pi(\theta) = 1$ leads to $\theta \,|\, \mathcal{D}_n \sim \mathrm{Beta}(S_n + 1, n - S_n + 1)$ as we saw earlier, which seemed very reasonable. But unfettered use of flat priors raises some questions.

Improper Priors. Let $X \sim N(\theta, \sigma^2)$ with σ known. We denote $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ as the observed data. Suppose we adopt a flat prior $\pi(\theta) \propto c$ where $c > 0$ is a constant. Note that $\displaystyle\int \pi(\theta) d\theta = \infty$ so this is not a valid probability density. We call such a prior an improper prior. Nonetheless, we can still formally carry out Bayes' theorem and compute the posterior density by multiplying the prior and the likelihood:

$$p(\theta \,|\, \mathcal{D}_n) \propto \mathcal{L}_n(\theta)\pi(\theta) \propto \mathcal{L}_n(\theta).$$

Let $\overline{X} = \sum_{i=1}^n X_i/n$. This gives $\theta \,|\, \mathcal{D}_n \sim N(\overline{X}, \sigma^2/n)$ and the resulting Bayesian point and interval estimators agree exactly with their frequentist counterparts. In general, improper priors are not a problem as long as the resulting posterior is a well-defined probability distribution.

Flat Priors are Not Invariant. Let $X \sim \mathrm{Bernoulli}(\theta)$ and suppose we use the flat prior $\pi(\theta) = 1$. This flat prior presumably represents our lack of information about θ before the experiment. Now let $\psi = \log(\theta/(1-\theta))$. This is a transformation of θ and we can compute the resulting distribution for ψ, namely,

$$p(\psi) = \frac{e^\psi}{(1 + e^\psi)^2},$$

which is not flat. But if we are ignorant about θ then we are also ignorant about ψ so we should use a flat prior for ψ. This is a contradiction! In short, the notion of a flat prior is not well defined because a flat prior on a parameter does not imply a flat prior on a transformed version of this parameter. Flat priors are not transformation invariant.

Jeffreys' Prior. To define priors that are transformation invariant, Harold Jeffreys came up with a rule: take the prior distribution on parameter space that is proportional to the square root of the determinant of the Fisher information.

$$\pi(\theta) \propto \sqrt{|I(\theta)|} \;\text{ where } I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log p(X \,|\, \theta)}{\partial\theta\partial\theta^T} \,\middle|\, \theta\right]$$

is the Fisher information.

There are various reasons for thinking that this prior might be a useful prior but we will not go into details here. The next theorem shows its transformation invariant property.

Theorem 210. The Jeffreys' prior is transformation invariant.

Proof. Let the likelihood function be $p(\mathrm{x}\,|\,\theta)$ and ψ be a transformation of θ, we need to show that $\pi(\psi) \propto \sqrt{|I(\psi)|}$. This result follows from the change of variable theorem and the fact that the product of determinants is the determinant of matrix product. □

Example 211. Consider the Bernoulli(θ) model. Recall that

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

Jeffreys' rule uses the prior

$$\pi(\theta) \propto \sqrt{I(\theta)} = \theta^{-1/2}(1-\theta)^{-1/2}.$$

This is a Beta (1/2,1/2) density and is very close to a uniform density.

The Jeffreys' prior is transformation invariant but this does not mean it is "noninformative". Researchers have tried to develop more sophisticated noninformative priors like reference priors [9, 7]. The reference prior coincides with the Jeffrey's prior for single-parameter models. For general multiparameter models, they can be different.

12.2.6 Conjugate Priors

We have already seen examples of conjugate priors above, with the binomial/Beta and multinomial/Dirichlet families. Here we first look at conjugacy from a more general perspective, and then give further examples.

Loosely speaking, a prior distribution is conjugate if it is closed under sampling. That is, suppose that $\mathcal{P}$ is a family of prior distributions, and for each θ, we have a distribution $p(\cdot\,|\,\theta) \in \mathcal{F}$ over a sample space $\mathcal{X}$. Then if the posterior

$$p(\theta\,|\,\mathrm{x}) = \frac{p(\mathrm{x}\,|\,\theta)\,\pi(\theta)}{\displaystyle\int p(\mathrm{x}\,|\,\theta)\,\pi(\theta)\,d\theta}$$

satisfies $p(\cdot\,|\,\mathrm{x}) \in \mathcal{P}$, we say that the family $\mathcal{P}$ is conjugate to the family of sampling distributions $\mathcal{F}$. In order for this to be a meaningful notion, the family $\mathcal{P}$ should be sufficiently restricted, and is typically taken to be a specific parametric family.

We can characterize the conjugate priors for general exponential family models. Suppose that $p(\cdot \,|\, \boldsymbol{\theta})$ is a standard exponential family model, where the densities with respect to a positive measure μ take the form

$$p(\mathbf{x} \,|\, \boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^T \mathbf{x} - A(\boldsymbol{\theta})) \tag{12.5}$$

where the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ is d-dimensional, and the parameter space $\Theta \subset \mathbb{R}^d$ is open, with

$$\int \exp\left(\boldsymbol{\theta}^T \mathbf{x} - A(\boldsymbol{\theta})\right) \, d\mu(\mathbf{x}) < \infty.$$

The moment generating function, or log-normalizing constant $A(\boldsymbol{\theta})$ is given by

$$A(\boldsymbol{\theta}) = \log \int \exp(\boldsymbol{\theta}^T \mathbf{x} - A(\boldsymbol{\theta})) \, d\mu(\mathbf{x}).$$

A conjugate prior for the exponential family (12.5) is a density of the form

$$\pi_{\mathbf{x}_0, n_0}(\theta) = \frac{\exp\left(n_0 \mathbf{x}_0^T \boldsymbol{\theta} - n_0 A(\boldsymbol{\theta})\right)}{\displaystyle\int \exp\left(n_0 \mathbf{x}_0^T \boldsymbol{\theta} - n_0 A(\boldsymbol{\theta})\right) d\boldsymbol{\theta}}$$

where $\mathbf{x}_0 \in \mathbb{R}^d$ is a vector and $n_0 \in \mathbb{R}$ is a scalar.

To see that this is conjugate, note that

$$\begin{aligned}
p(\mathbf{x} \,|\, \boldsymbol{\theta}) \, \pi_{\mathbf{x}_0, n_0}(\boldsymbol{\theta}) &= \exp\left(\boldsymbol{\theta}^T \mathbf{x} - A(\boldsymbol{\theta})\right) \exp\left(n_0 \mathbf{x}_0^T \boldsymbol{\theta} - n_0 A(\boldsymbol{\theta})\right) \\
&= \exp\left((\mathbf{x} + \mathbf{x}_0)^T \boldsymbol{\theta} - (1 + n_0) A(\boldsymbol{\theta})\right) \\
&= \exp\left((1 + n_0)\left(\frac{\mathbf{x}}{1 + n_0} + \frac{n_0 \mathbf{x}_0}{1 + n_0}\right)^T \boldsymbol{\theta} - (1 + n_0) A(\boldsymbol{\theta})\right) \\
&\propto \pi_{\frac{\mathbf{x}}{1+n_0} + \frac{n_0 \mathbf{x}_0}{1+n_0}, 1+n_0}(\boldsymbol{\theta}).
\end{aligned}$$

We can think of the prior as incorporating n_0 "virtual" observations of $\mathbf{x}_0 \in \mathbb{R}^d$. The parameters of the posterior after making one "real" observation x are then $n_0' = 1 + n_0$ and

$$\mathbf{x}_0' = \frac{\mathbf{x}}{1 + n_0} + \frac{n_0 \mathbf{x}_0}{1 + n_0}$$

which is a mixture of the virtual and actual observations. More generally, if we have n observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, then the posterior takes the form

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \pi_{\frac{n\bar{\mathbf{X}}}{n+n_0} + \frac{n_0 \mathbf{x}_0}{n+n_0}, n+n_0}(\boldsymbol{\theta})$$

$$\propto \quad \exp\Bigg((n+n_0) \left(\frac{n\overline{\boldsymbol{X}}}{n+n_0} + \frac{n_0 \mathbf{x}_0}{n+n_0} \right)^T \boldsymbol{\theta} - (n+n_0)A(\boldsymbol{\theta}) \Bigg),$$

where $\overline{\boldsymbol{X}} = \sum_{i=1}^n \boldsymbol{X}_i/n$. Thus, the parameters of the posterior are $n_0' = n + n_0$ and the mixture

$$\mathbf{x}_0' \quad = \quad \frac{n\overline{\boldsymbol{X}}}{n+n_0} + \frac{n_0 \mathbf{x}_0}{n+n_0}.$$

Now, let $\pi_{\mathbf{x}_0,n_0}$ be defined by

$$\pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta}) = \exp\left(n_0 \mathbf{x}_0^T \boldsymbol{\theta} - n_0 A(\boldsymbol{\theta}) \right),$$

so that

$$\nabla \pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta}) = n_0 \left(\mathbf{x}_0 - \nabla A(\boldsymbol{\theta}) \right) \pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta}).$$

Since

$$\int \nabla \pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \nabla \left(\int \pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right) = 0,$$

from which it follows that

$$\mathbb{E}[\nabla A(\boldsymbol{\theta})] = \int \nabla A(\boldsymbol{\theta}) \pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbf{x}_0 - \frac{1}{n_0} \int \nabla \pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathbf{x}_0,$$

where the expectation is with respect to $\pi_{\mathbf{x}_0,n_0}(\boldsymbol{\theta})$. More generally,

$$\mathbb{E}\left[\nabla A(\boldsymbol{\theta}) \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \right] = \frac{n\overline{\boldsymbol{X}}}{n_0+n} + \frac{n_0 \mathbf{x}_0}{n_0+n}.$$

Under appropriate regularity conditions, the converse also holds, so that linearity of

$$\mathbb{E}(\nabla A(\boldsymbol{\theta}) \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$$

is sufficient for conjugacy; this is the following result of Diaconis (1979), stated here in the continuous case.

Theorem 212. Suppose that $\Theta \subset \mathbb{R}^d$ is open, and let $\boldsymbol{X}$ be a sample of size one from the exponential family $p(\cdot \,|\, \boldsymbol{\theta})$, where the support of μ contains an open interval. Suppose that $\boldsymbol{\theta}$ has a prior density $\pi(\boldsymbol{\theta})$ which does not concentrate at a single point. Then the posterior mean of $\nabla A(\boldsymbol{\theta})$ given a single observation $\boldsymbol{X}$ is linear,

$$\mathbb{E}(\nabla A(\theta) \,|\, X) = a\boldsymbol{X} + \boldsymbol{b},$$

if and only if the prior $\pi(\boldsymbol{\theta})$ is given by

$$\pi(\boldsymbol{\theta}) \propto \exp\left(\frac{1}{a} \boldsymbol{b}^T \boldsymbol{\theta} - \frac{1-a}{a} A(\boldsymbol{\theta}) \right).$$

A similar result holds in the case where μ is a discrete measure, as in the case of the multinomial family.

First consider the Poisson model with rate $\lambda \geq 0$, given by sample space $\mathcal{X} = \mathbb{Z}_+$ and

$$\mathbb{P}(X = \mathrm{x} \,|\, \lambda) = \frac{\lambda^{\mathrm{x}}}{\mathrm{x}!} e^{-\lambda} \propto \exp(\mathrm{x} \log \lambda - \lambda).$$

Thus the natural parameter is $\theta = \log \lambda$, and the conjugate prior takes the form

$$\pi_{\mathrm{x}_0, n_0}(\lambda) \propto \exp(n_0 \mathrm{x}_0 \log \lambda - n_0 \lambda).$$

A better parameterization of the prior is

$$\pi_{\alpha,\beta}(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$$

which is the Gamma(α, β) density. Using this parameterization, let $X_1, \ldots, X_n$ be observations from Poisson(λ), we see that the posterior is given by

$$\lambda \,|\, X_1 \ldots, X_n \quad \sim \quad \text{Gamma}(\alpha + n\overline{X}, \beta + n).$$

Here we see that the prior acts as if β virtual observations were made, with a total count of $\alpha - 1$ among them.

Next consider the *exponential model*, where the sample space $\mathcal{X} = \mathbb{R}_+$ is the non-negative real line, and

$$p(\mathrm{x} \,|\, \theta) = \theta e^{-\mathrm{x}\theta}.$$

This is a widely used model for survival times or waiting times between events. The conjugate prior, in the most convenient parameterization, is again the Gamma

$$\pi_{\alpha,\beta}(\theta) \quad \propto \quad \theta^{\alpha-1} e^{-\beta\theta}.$$

Let $X_1, \ldots, X_n$ be observed data from Gamma(θ). The posterior is given by

$$\theta \,|\, X_1, \ldots, X_n \quad \sim \quad \text{Gamma}(\alpha + n, \beta + n\overline{X}).$$

Thus, in this case the prior acts as if $\alpha - 1$ virtual examples are used, with a total waiting time of β.

The discrete analogue of the exponential model is the geometric distribution, with sample space $\mathcal{X} = \mathbb{Z}_{++}$ the strictly positive integers and sampling distribution

$$\mathbb{P}(X = \mathrm{x} \,|\, \theta) = (1 - \theta)^{\mathrm{x}-1} \theta.$$

The conjugate prior for this model is the Gamma(α, β) distribution. Let $X_1, \ldots, X_n$ be observed data from Geometric(θ). The posterior is

$$\theta \,|\, X_1, \ldots, X_n \quad \sim \quad \text{Gamma}(\alpha + n, \beta + n\overline{X})$$

just as for the exponential model.

Next, consider a Gaussian model with known mean μ, so that the free parameter is the variance σ^2. The likelihood function is

$$\begin{aligned} p(X_1,\ldots,X_n \,|\, \sigma^2) &\propto (\sigma^2)^{-n/2}\exp\biggl(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\biggr) \\ &= (\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}n\overline{(X-\mu)^2}\right) \end{aligned}$$

where

$$\overline{(X-\mu)^2} = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2.$$

The conjugate prior is an inverse Gamma distribution. Recall that θ has an inverse Gamma distribution with parameters α and β in case $1/\theta \sim \text{Gamma}(\alpha,\beta)$; the density takes the form

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{-(\alpha+1)}e^{-\beta/\theta}.$$

With this prior, the posterior distribution of σ^2 is given by

$$\sigma^2 \,|\, X_1,\ldots,X_n \sim \text{InvGamma}\left(\alpha+\frac{n}{2}, \beta+\frac{n}{2}\overline{(X-\mu)^2}\right).$$

Alternatively, the prior can be parameterized in terms of the scaled inverse χ^2 distribution, which has density of the form

$$\pi_{\nu_0,\sigma_0^2}(\theta) \propto \theta^{-(1+\nu_0/2)}\exp\left(-\frac{\nu_0\sigma_0^2}{2\theta}\right).$$

This is the distribution of $\sigma_0^2\nu_0 Z$ where $Z \sim \chi^2_{\nu_0}$. Under this prior, the posterior takes the form

$$\sigma^2 \,|\, X_1,\ldots,X_n \sim \text{ScaledInv-}\chi^2\biggl(\nu_0+n, \frac{\nu_0\sigma_0^2}{\nu_0+n} + \frac{n\overline{(X-\mu)^2}}{\nu_0+n}\biggr).$$

In the multidimensional setting, the inverse Wishart takes the place of the inverse Gamma. The Wishart distribution is a multidimensional analogue of the Gamma distribution; it is a distribution over symmetric positive semi-definite $d \times d$ matrices $\mathbf{W}$ with density of the form

$$\pi_{\nu_0,\mathbf{S}_0}(\mathbf{W}) \propto |\mathbf{W}|^{(\nu_0-d-1)/2}\exp\left(-\frac{1}{2}\text{tr}(\mathbf{S}_0^{-1}\mathbf{W})\right),$$

where the parameters are the degrees of freedom ν_0 and the positive-definite matrix $\mathbf{S}_0$. If $\mathbf{W}^{-1} \sim$ Wishart$(\nu_0, \mathbf{S}_0)$ then $\mathbf{W}$ has an inverse Wishart distribution; the density of the inverse Wishart has the form

$$\pi_{\nu_0,\mathbf{S}_0}(\mathbf{W}) \propto |\mathbf{W}|^{-(\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{S}_0\mathbf{W}^{-1})\right).$$

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be observed data from $N(\mathbf{0}, \boldsymbol{\Sigma})$, then an inverse Wishart prior multiplies the likelihood takes the form

$$\begin{aligned} & p(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \,|\, \boldsymbol{\Sigma})\, \pi_{\nu_0,\mathbf{S}_0}(\boldsymbol{\Sigma}) \propto \\ & \quad |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{n}{2}\text{tr}(\overline{\mathbf{S}}\boldsymbol{\Sigma}^{-1})\right) |\boldsymbol{\Sigma}|^{-(\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{S}_0\boldsymbol{\Sigma}^{-1})\right) \\ & = |\boldsymbol{\Sigma}|^{-(n+\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\text{tr}((n\overline{\mathbf{S}} + \mathbf{S}_0)\boldsymbol{\Sigma}^{-1})\right) \end{aligned}$$

where $\overline{\mathbf{S}}$ is the empirical covariance $\overline{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T$. Thus, the posterior takes the form

$$\boldsymbol{\Sigma} \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \text{InvWishart}(\nu_0 + n, \mathbf{S}_0 + n\overline{\mathbf{S}}).$$

Similarly, the conjugate prior for the inverse covariance $\boldsymbol{\Sigma}^{-1}$ (precision matrix) is a Wishart.

The previous examples are all for exponential family distributions. Now we consider a non-exponential family example, the uniform distribution Unifom$(0, \theta)$ with parameter $\theta \geq 0$. Recall that the Pareto is the standard power-law distribution, if $\theta \sim$ Pareto(ν_0, k), the survival function is

$$\mathbb{P}(\theta \geq t) = \left(\frac{t}{\nu_0}\right)^{-k}, \quad t \geq \nu_0.$$

The parameter k determines the rate of decay, and ν_0 specifies the support of the distribution. The density is given by

$$\pi_{k,\nu_0}(\theta) = \begin{cases} \dfrac{k\nu_0^k}{\theta^{k+1}} & \theta \geq \nu_0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $X_1, \ldots, X_n$ be observed data from Uniform$(0, \theta)$ and the prior of θ is Pareto(k, ν_0). Let $X_{(n)} = \max_{1\leq i\leq n}\{X_i\}$. If $\nu_0 > X_{(n)}$, then $\mathcal{L}_n(\theta)\pi_{k,\nu_0}(\theta) = 0$. But if $X_{(n)} \geq \nu_0$, then under the posterior we know that θ must be at least $X_{(n)}$, and in this case

$$\mathcal{L}_n(\theta)\pi_{k,\nu_0}(\theta) \propto \frac{1}{\theta^n}\frac{1}{\theta^{k+1}}.$$

Thus, the posterior is given by

$$\theta \,|\, X_1, \ldots, X_n \sim \text{Pareto}\left(n + k, \max\{X_{(n)}, \nu_0\}\right).$$

Thus, as n increases, the decay of the posterior increases, resulting in a more peaked distribution around $X_{(n)}$; the parameter k controls the sharpness of the decay for small n.

Charts with models and conjugate priors for both discrete and continuous distributions are provided in Figures 12.3 and 12.4.

Sample Space	Sampling Dist.	Conjugate Prior	Posterior
$\mathcal{X} = \{0, 1\}$	Bernoulli(θ)	Beta(α, β)	Beta($\alpha + n\overline{X}, \beta + n(1 - \overline{X})$)
$\mathcal{X} = \mathbb{Z}_+$	Poisson(λ)	Gamma(α, β)	Gamma($\alpha + n\overline{X}, \beta + n$)
$\mathcal{X} = \mathbb{Z}_{++}$	Geometric(θ)	Gamma(α, β)	Gamma($\alpha + n, \beta + n\overline{X}$)
$\mathcal{X} = \mathbb{H}_K$	Multinomial(θ)	Dirichlet(α)	Dirichlet($\alpha + n\overline{X}$)

Figure 12.3: Conjugate priors for discrete exponential family distributions.

12.2.7 Bayesian Hypothesis Testing

Suppose we want to test the following hypothesis:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

where $\theta \in \mathbb{R}$. The Bayesian approach to testing involves putting a prior on H_0 and on the parameter θ and then computing $\mathbb{P}(H_0 \,|\, \mathcal{D}_n)$. It is common to use the prior $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ (although this is not essential in what follows). Under H_1 we need a prior for θ. Denote this prior density by $\pi(\theta)$. In such a setting, the prior distribution comprise a point mass 0.5 at θ_0 mixed with a continuous density elsewhere. From Bayes' theorem

$$\begin{aligned}
\mathbb{P}(H_0 \,|\, \mathcal{D}_n) &= \frac{p(\mathcal{D}_n \,|\, H_0)\mathbb{P}(H_0)}{p(\mathcal{D}_n \,|\, H_0)\mathbb{P}(H_0) + p(\mathcal{D}_n \,|\, H_1)\mathbb{P}(H_1)} \\
&= \frac{p(\mathcal{D}_n \,|\ \ \theta_0)}{p(\mathcal{D}_n \,|\ \ \theta_0) + p(\mathcal{D}_n \,|\ \ H_1)} \\
&= \frac{p(\mathcal{D}_n \,|\ \ \theta_0)}{p(\mathcal{D}_n \,|\ \ \theta_0) + \displaystyle\int p(\mathcal{D}_n \,|\ \ \theta)\pi(\theta)d\theta}
\end{aligned}$$

Sampling Dist.	Conjugate Prior	Posterior
Uniform(θ)	Pareto(ν_0, k)	Pareto $\big(\max\{\nu_0, X_{(n)}\}, n+k\big)$
Exponential(θ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + n\overline{X})$
$N(\mu, \sigma^2),$ known σ^2	$N(\mu_0, \sigma_0^2)$	$N\left(\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{X}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$
$N(\mu, \sigma^2),$ known μ	InvGamma(α, β)	InvGamma $\left(\alpha + \frac{n}{2}, \beta + \frac{n}{2}\overline{(X-\mu)^2}\right)$
$N(\mu, \sigma^2),$ known μ	ScaledInv-$\chi^2(\nu_0, \sigma_0^2)$	ScaledInv-χ^2 $\left(\nu_0 + n, \frac{\nu_0\sigma_0^2}{\nu_0 + n} + \frac{n\overline{(X-\mu)^2}}{\nu_0 + n}\right)$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$ known $\boldsymbol{\Sigma}$	$N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$	$N\left(\mathbf{K}\left(\boldsymbol{\Sigma}_0^{-1}\mu_0 + n\boldsymbol{\Sigma}^{-1}\overline{X}\right), \mathbf{K}\right),\ \mathbf{K} = \left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$ known $\boldsymbol{\mu}$	InvWishart$(\nu_0, \mathbf{S}_0)$	InvWishart$(\nu_0 + n, \mathbf{S}_0 + n\overline{\mathbf{S}})$, $\overline{\mathbf{S}}$ sample covariance

Figure 12.4: Conjugate priors for some continuous distributions.

$$= \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \int \mathcal{L}(\theta)\pi(\theta)d\theta}.$$

We saw that, in estimation problems, the prior was not very influential and that the frequentist and Bayesian methods gave similar answers. This is not the case in hypothesis testing. Also, one can't use improper priors in testing because this leads to an undefined constant in the denominator of the expression above. Thus, if you use Bayesian testing you must choose the prior $\pi(\theta)$ very carefully.

12.2.8 Model Comparison and Bayesian Information Criterion

Let $\mathcal{D}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be the data. Suppose we consider K parametric models $\mathcal{M}_1, \ldots, \mathcal{M}_K$. In Bayesian inference, we assign a prior probability $\pi_j = \mathbb{P}(\mathcal{M}_j)$ to model $\mathcal{M}_j$ and prior $p_j(\boldsymbol{\theta}_j \mid \mathcal{M}_j)$ to the parameters $\boldsymbol{\theta}_j$ under model $\mathcal{M}_j$. The posterior probability of model $\mathcal{M}_j$ conditional on data $\mathcal{D}_n$ is

$$\mathbb{P}(\mathcal{M}_j \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \mathcal{M}_j)\pi_j}{p(\mathcal{D}_n)} = \frac{p(\mathcal{D}_n \mid \mathcal{M}_j)\pi_j}{\sum_{k=1}^K p(\mathcal{D}_n \mid \mathcal{M}_k)\pi_k}$$

where

$$p(\mathcal{D}_n \,|\, \mathcal{M}_j) = \int \mathcal{L}_j(\theta_j) p_j(\theta_j) d\theta_j$$

and $\mathcal{L}_j$ is the likelihood function for model j. Hence,

$$\frac{\mathbb{P}(\mathcal{M}_j \,|\, \mathcal{D}_n)}{\mathbb{P}(\mathcal{M}_k \,|\, \mathcal{D}_n)} = \frac{p(\mathcal{D}_n \,|\, \mathcal{M}_j)\pi_j}{p(\mathcal{D}_n \,|\, \mathcal{M}_k)\pi_k}. \tag{12.6}$$

To choose between models $\mathcal{M}_j$ and $\mathcal{M}_k$, we examine the right hand side of (12.6). If it's larger than 1, we prefer model $\mathcal{M}_j$; otherwise, we prefer model $\mathcal{M}_k$.

Definition 213. (Bayes factor) The Bayes factor between models $\mathcal{M}_j$ and $\mathcal{M}_k$ is defined to be

$$\mathrm{BF}(\mathcal{D}_n) = \frac{p(\mathcal{D}_n \,|\, \mathcal{M}_j)}{p(\mathcal{D}_n \,|\, \mathcal{M}_k)} = \frac{\int \mathcal{L}_j(\theta_j) p_j(\theta_j) d\theta_j}{\int \mathcal{L}_k(\theta_k) p_k(\theta_k) d\theta_k}.$$

Here, $p(\mathcal{D}_n \,|\, \mathcal{M}_j)$ is called the marginal likelihood for model $\mathcal{M}_j$.

The use of Bayes factors can be viewed as a Bayesian alternative to classical hypothesis testing. Bayesian model comparison is a method of model selection based on Bayes factors.

In a typical setting, we adopt a uniform prior over the models: $\pi_1 = \pi_2 = \ldots = \pi_K = 1/K$. Now

$$p(\mathcal{D}_n \,|\, \mathcal{M}_j) = \int p(\mathcal{D}_n \,|\, \mathcal{M}_j, \boldsymbol{\theta}_j) p_j(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j) d\boldsymbol{\theta}_j = \int \mathcal{L}_n(\boldsymbol{\theta}_j) p_j(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j) d\boldsymbol{\theta}_j.$$

Under model $\mathcal{M}_j$, we define $I_n(\boldsymbol{\theta}_j)$ and $I_1(\boldsymbol{\theta}_j)$ to be the empirical Fisher information matrices for the dataset $\mathcal{D}_n$ and one data point:

$$I_n(\boldsymbol{\theta}_j) = -\frac{\partial^2 \log p(\mathcal{D}_n \,|\, \mathcal{M}_j, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_j^T} \text{ and } I_1(\boldsymbol{\theta}_j) = -\frac{\partial^2 \log p(\boldsymbol{X}_1 \,|\, \mathcal{M}_j, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_j^T}.$$

Recall that, under certain regularity conditions, $I_n(\theta_j) = nI_1(\theta_j)$. Let $\widehat{\boldsymbol{\theta}}_j$ be the maximum a posterior (MAP) estimator under model $\mathcal{M}_j$, i.e.,

$$\left. \frac{\partial \log p(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j, \mathcal{D}_n)}{\partial \boldsymbol{\theta}_j} \right|_{\boldsymbol{\theta}_j = \widehat{\boldsymbol{\theta}}_j} = 0.$$

Let $\mathcal{L}_n(\boldsymbol{\theta}_j) = p(\mathcal{D}_n \,|\, \mathcal{M}_j, \boldsymbol{\theta}_j)$. By a Taylor expansion at $\widehat{\boldsymbol{\theta}}_j$, we have

$$\log \mathcal{L}_n(\boldsymbol{\theta}_j) \approx \log \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) - \frac{1}{2}(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j)^T I_n(\widehat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j).$$

Therefore, by exponentiating both sides,

$$\mathcal{L}_n(\boldsymbol{\theta}_j) \approx \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j)^T I_n(\widehat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j)\right).$$

Assuming $\boldsymbol{\theta}_j \in \mathbb{R}^{d_j}$, we choose a prior $p_j(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j)$ that is noninformative or "flat" over the neighborhood of $\widehat{\boldsymbol{\theta}}_k$ where $\mathcal{L}_n(\boldsymbol{\theta})$ is dominant. We then have

$$\begin{aligned}
& p(\mathcal{D}_n \,|\, \mathcal{M}_j) \\
&= \int \mathcal{L}_n(\boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j) d\boldsymbol{\theta}_j \\
&= \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) \int \frac{\mathcal{L}_n(\boldsymbol{\theta}_j)}{\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j)} p_j(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j) d\boldsymbol{\theta}_j \\
&\approx \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) p_j(\widehat{\boldsymbol{\theta}}_j \,|\, \mathcal{M}_j) \int \frac{\mathcal{L}_n(\boldsymbol{\theta}_j)}{\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j)} d\boldsymbol{\theta}_j \\
&\approx \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) p_j(\widehat{\boldsymbol{\theta}}_j \,|\, \mathcal{M}_j) \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j)^T I_n(\widehat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j)\right\} d\boldsymbol{\theta}_j \\
&= \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) p_j(\widehat{\boldsymbol{\theta}}_j \,|\, \mathcal{M}_j) \frac{(2\pi)^{d_j/2}}{|I_n(\widehat{\boldsymbol{\theta}}_j)|^{1/2}} \qquad (12.7) \\
&= \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) p(\widehat{\boldsymbol{\theta}}_j \,|\, \mathcal{M}_j) \frac{(2\pi)^{d_j/2}}{n^{d_j/2}|I_1(\widehat{\boldsymbol{\theta}}_j)|^{1/2}}.
\end{aligned}$$

Equation (12.7) was obtained by recognizing that the integrand is the kernel of a Gaussian density.

Now

$$-2\log p(\mathcal{D}_n \,|\, \mathcal{M}_j) \approx -2\log \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) + d_j \log n + \log|I_1(\widehat{\boldsymbol{\theta}}_j)| - d_j \log(2\pi) + \log p(\widehat{\boldsymbol{\theta}}_j \,|\, \mathcal{M}_j).$$

The term $\log|I_1(\widehat{\boldsymbol{\theta}}_j)| - d_j \log(2\pi) + \log p(\widehat{\boldsymbol{\theta}}_j \,|\, \mathcal{M}_j)$ is of smaller order than $-2\log \mathcal{L}_n(\widehat{\boldsymbol{\theta}}_j) + d_j \log n$. Hence, we can approximate $\log p(\mathcal{D}_n \,|\, \mathcal{M}_j)$ with Bayesian information criterion BIC) defined as:

Definition 214. (Bayesian information criterion) Given data $\mathcal{D}_n$ and a model $\mathcal{M}$, the Bayesian information criterion for $\mathcal{M}$ is defined to be

$$\text{BIC}(\mathcal{M}) = \log \mathcal{L}_j(\theta_j) - \frac{d}{2}\log n,$$

where d is the dimensionality of the of model $\mathcal{M}_j$.

The BIC score provides a large-sample approximation to the log posterior probability associated with the approximating model. By choosing the fitted candidate model corresponding to the maxium value of BIC, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability.

It is easy to see that

$$\log \frac{p(\mathcal{D}_n \mid \mathcal{M}_j)}{p(\mathcal{D}_n \mid \mathcal{M}_k)} = \mathrm{BIC}(\mathcal{M}_j) - \mathrm{BIC}(\mathcal{M}_k) + O_P(1).$$

This relationship implies that, if $\pi_1 = \ldots = \pi_K = 1/K$, then

$$p(\mathcal{M}_j \mid \mathcal{D}_n) \approx \frac{\exp\big(\mathrm{BIC}(\mathcal{M}_j)\big)}{\sum_{k=1}^{K} \exp\left(\mathrm{BIC}(\mathcal{M}_k)\right)}.$$

Typically, $\log(p(\mathcal{D}_n \mid \mathcal{M}_j)/p(\mathcal{D}_n \mid \mathcal{M}_k))$ tends to ∞ or $-\infty$ as $n \to \infty$ in which case the $O_P(1)$ term is negligible. This justifies BIC as an approximation to the posterior. More precise approximations are possible by way of simulation. However, the improvements are limited to the $O_P(1)$ error term. Compared to AIC, BIC prefers simpler models. In fact, we can show that BIC is model selection consistent, i.e. if the true model is within the candidate pool, the probability that BIC selects the true model goes to 1 as n goes to infinity.

However, BIC does not select the fitted candidate model which minimizes the mean squared error for prediction. In contrast, AIC does optomize predictive accuracy.

12.2.9 Calculating the Posterior Distribution

To compute any marginal of the posterior distribution $p(\theta \mid \mathcal{D}_n)$ usually involves high dimensional integration. Uusally, we instead approximate the marginals by simulation methods.

Suppose we draw $\theta^1, \ldots, \theta^B \sim p(\theta \mid \mathcal{D}_n)$. Then a histogram of $\theta^1, \ldots, \theta^B$ approximates the posterior density $p(\theta \mid \mathcal{D}_n)$. An approximation to the posterior mean $\overline{\theta}_n = \mathbb{E}(\theta \mid \mathcal{D}_n)$ is $B^{-1} \sum_{j=1}^{B} \theta^j$. The posterior $1 - \alpha$ interval can be approximated by $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$ where $\theta_{\alpha/2}$ is the $\alpha/2$ sample quantile of $\theta^1, \ldots, \theta^B$. Once we have a sample $\theta^1, \ldots, \theta^B$ from $p(\theta \mid \mathcal{D}_n)$, let $\tau^i = g(\theta^i)$. Then $\tau^1, \ldots, \tau^B$ is a sample from $p(\tau \mid \mathcal{D}_n)$. This avoids the need to do any integration.

In this section, we will describe methods for obtaining simulated values from the posterior. The simulation methods we discuss include Monte Carlo integration, importance sampling, and Markov chain Monte Carlo (MCMC). We will also describe another approximation method called variational inference. While variational methods and stochastic simulation methods such as MCMC address many of the same problems, they differ greatly in their approach. Variational methods are based on deterministic approximation and numerical optimization, while simulation methods are based on random sampling. Variational methods have been successfully applied to a wide range of problems, but they come with very weak theoretical guarantees.

Example 215. Consider again Example 208. We can approximate the posterior for ψ without doing any calculus. Here are the steps:

1. Draw $\theta^1, \ldots, \theta^B \sim \mathrm{Beta}(s+1, n-s+1)$.
2. Let $\psi^i = \log(\theta^i/(1-\theta^i))$ for $i = 1, \ldots, B$.

Now $\psi^1, \ldots, \psi^B$ are i.i.d. draws from the posterior density $p(\psi \,|\, \mathcal{D}_n)$. A histogram of these values provides an estimate of $p(\psi \,|\, \mathcal{D}_n)$.

12.3 Theoretical Aspects of Bayesian Inference

In this section we explain some theory related to the Bayesian inference. In particular, we discuss the frequentist aspects of Bayesian procedures.

12.3.1 Bayesian Decision Theory

Let $\widehat{\theta}(X)$ be an estimator of a parameter $\theta \in \Theta$. The notation $\widehat{\theta}(X)$ reflects the fact that $\widehat{\theta}$ is a function of the data X. We measure the discrepancy between a parameter θ and its estimator $\widehat{\theta}(X)$ using a loss function $L : \Theta \times \Theta \to \mathbb{R}$. We define the risk of an estimator $\widehat{\theta}(X)$ as

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\big(L(\theta, \widehat{\theta})\big) = \int L(\theta, \widehat{\theta}(\mathrm{x}))\, p_\theta(\mathrm{x})\, d\mathrm{x}.$$

From a frequentist viewpoint, the parameter θ is a deterministic quantity. In frequentist inference we ofetn try to find a minimax estimator $\widehat{\theta}$ which is an estimator that minimizes the maximum risk

$$R_{\max}(\widetilde{\theta}) := \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}).$$

From a Bayesian viewpoint, the parameter θ is a random quantity with a prior distribution $\pi(\theta)$. The Bayesian approach to decision theory is to find the estimator $\widehat{\theta}(X)$ that minimizes the posterior expected loss

$$R_\pi(\widehat{\theta}|X) = \int_\Theta L(\theta, \widehat{\theta}(X)) p(\theta \,|\, X) d\theta.$$

An estimator $\widehat{\theta}$ is a Bayes rule with respect to the prior $\pi(\theta)$ if

$$R_\pi(\widehat{\theta}) = \inf_{\widetilde{\theta} \in \Theta} R_\pi(\widetilde{\theta}|X),$$

where the infimum is over all estimators $\widetilde{\theta} \in \Theta$.

It turns out that minimizing the posterior expected loss is equivalent to minimizing the average risk, also known as the Bayes risk defined by

$$B_\pi = \int R(\theta, \widehat{\theta})\pi(\theta) d\theta.$$

Theorem 216. The Bayes rule minimizes the B_π.

Under different loss functions, we get different estimators.

Theorem 217. If $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ then the Bayes estimator is the posterior mean. If $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$ then the Bayes estimator is the posterior median. If θ is discrete and $L(\theta, \widehat{\theta}) = I(\theta \neq \widehat{\theta})$ then the Bayes estimator is the posterior mode.

12.3.2 Large Sample Properties of Bayes' Procedures

Under appropriate conditions, the posterior distribution tends to a Normal distribution. Also, the posterior mean and the mle are very close. The proof of the following theorem can be found in the van der Vaart (1998).

Theorem 218. Let $I(\theta)$ denote the Fisher information. Let $\widehat{\theta}_n$ be the maximum likelihood estimator and let

$$\widehat{\text{se}} = \frac{1}{\sqrt{nI(\widehat{\theta}_n)}}.$$

Under appropriate regularity conditions, the posterior is approximately Normal with mean $\widehat{\theta}_n$ and standard deviation $\widehat{\text{se}}$. That is,

$$\int |p(\theta|X_1, \ldots, X_n) - \phi(\theta; \widehat{\theta}_n, \text{se})| d\theta \xrightarrow{P} 0.$$

Also, $\overline{\theta}_n - \widehat{\theta}_n = O_P(1/n)$. Let $z_{\alpha/2}$ be the $\alpha/2$-quantile of a standard Gaussian distribution and let $C_n = [\widehat{\theta}_n - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta}_n + z_{\alpha/2}\widehat{\text{se}}]$ be the asymptotic frequentist $1-\alpha$ confidence interval. Then

$$\mathbb{P}(\theta \in C_n \,|\, \mathcal{D}_n) \to 1 - \alpha.$$

Proof. Here we only give a proof outline. See Chapter 10 of van der Vaart (1998) for a rigorous proof. It can be shown that the effect of the prior diminishes as n increases so that $p(\theta \,|\, \mathcal{D}_n) \propto \mathcal{L}_n(\theta)p(\theta) \approx \mathcal{L}_n(\theta)$. Let $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$, we have $\log p(\theta \,|\, \mathcal{D}_n) \approx \ell_n(\theta)$. Now, by a Taylor expansion around $\widehat{\theta}_n$,

$$\begin{aligned}\ell_n(\theta) &\approx \ell_n(\widehat{\theta}_n) + (\theta - \widehat{\theta}_n)\ell_n'(\widehat{\theta}) + [(\theta - \widehat{\theta}_n)^2/2]\ell_n''(\widehat{\theta}_n) \\ &= \ell_n(\widehat{\theta}_n) + [(\theta - \widehat{\theta}_n)^2/2]\ell_n''(\widehat{\theta}_n),\end{aligned}$$

since $\ell_n'(\widehat{\theta}_n) = 0$. Exponentiating both sides, we get that, approximately,

$$p(\theta \,|\, \mathcal{D}_n) \propto \exp\Big\{-\frac{1}{2}\frac{(\theta - \widehat{\theta}_n)^2}{\sigma_n^2}\Big\},$$

where $\sigma_n^2 = -1/\ell''(\widehat{\theta}_n)$. So the posterior of θ is approximately Normal with mean $\widehat{\theta}_n$ and variance σ_n^2. Let $\ell_i(\theta) = \log p(X_i \,|\, \theta)$, then

$$\frac{1}{\sigma_n^2} \;=\; -\ell''(\widehat{\theta}_n) = \sum_{i=1}^n -\ell_i''(\widehat{\theta}_n) = n\left(\frac{1}{n}\right)\sum_{i=1}^n -\ell_i''(\widehat{\theta}_n) \approx n\mathbb{E}_\theta\left[-\ell_i''(\widehat{\theta}_n)\right] = nI(\widehat{\theta}_n)$$

and hence $\sigma_n \approx \mathrm{se}(\widehat{\theta}_n)$. □

There is also a Bayesian delta method. Let $\tau = g(\theta)$. Then $\tau \,|\, \mathcal{D}_n \approx N(\widehat{\tau}, \widetilde{\mathrm{se}}^2)$ where $\widehat{\tau} = g(\widehat{\theta})$ and $\widetilde{\mathrm{se}} = \widehat{\mathrm{se}}\,|g'(\widehat{\theta})|$.

12.4 Examples of Bayesian Inference

We now illustrate Bayesian inference with some examples.

12.4.1 Bayesian Linear Models

Many frequentist methods can be viewed as the maximum a posterior (MAP) estimator under a Bayesian framework. As an example, we consider Gaussian linear regression:

$$Y = \beta_0 + \sum_{j=1}^d \beta_j X_j + \epsilon, \;\; \epsilon \sim N(0, \sigma^2).$$

Here we assume that σ is known. Let $\mathcal{D}_n = \big\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\big\}$ be the observed data points. The conditional likelihood of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^n p(y_i \,|\, x_i, \boldsymbol{\beta}) \propto \exp\bigg(-\frac{\sum_{i=1}^n \Big(y_i - \beta_0 - \sum_{j=1}^d \beta_j x_{ij}\Big)^2}{2\sigma^2}\bigg).$$

Using a Gaussian prior $\pi_\lambda(\boldsymbol{\beta}) \propto \exp\big(-\lambda\|\boldsymbol{\beta}\|_2^2/2\big)$, the posterior of $\boldsymbol{\beta}$ can be written as

$$p(\boldsymbol{\beta}\,|\,\mathcal{D}_n) \propto \mathcal{L}(\boldsymbol{\beta})\pi_\lambda(\boldsymbol{\beta}).$$

The MAP estimator $\widehat{\boldsymbol{\beta}}^{\rm MAP}$ takes the form

$$\widehat{\boldsymbol{\beta}}^{\rm MAP} = \mathop{\rm argmax}_{\boldsymbol{\beta}} p(\boldsymbol{\beta}\,|\,\mathcal{D}_n) = \mathop{\rm argmin}_{\boldsymbol{\beta}}\bigg\{\sum_{i=1}^n\Big(Y_i - \beta_0 - \sum_{j=1}^d \beta_j X_{ij}\Big)^2 + \lambda\sigma^2\|\boldsymbol{\beta}\|_2^2\bigg\}.$$

This is exactly the ridge regression with the regularization parameter $\lambda' = \lambda\sigma^2$. If we adopt the Laplacian prior $\pi_\lambda(\boldsymbol{\beta}) \propto \exp\left(-\lambda\|\boldsymbol{\beta}\|_1/2\right)$, we get the Lasso estimator

$$\widehat{\boldsymbol{\beta}}^{\rm MAP} = \mathop{\rm argmin}_{\boldsymbol{\beta}}\bigg\{\sum_{i=1}^n\Big(Y_i - \beta_0 - \sum_{j=1}^d \beta_j X_{ij}\Big)^2 + \lambda\sigma^2\|\boldsymbol{\beta}\|_1\bigg\}.$$

Instead of using the MAP point estimate, a complete Bayesian analysis aims at obtaining the whole posterior distribution $p(\boldsymbol{\beta}\,|\,\mathcal{D}_n)$. In general, $p(\boldsymbol{\beta}\,|\,\mathcal{D}_n)$ does not have an analytic form and we need to resort to simulation to approximate the posterior.

12.4.2 Hierarchical Models

A hierarchical model is a multi-level statistical model that allows us to incorporate richer information into the model. A typical hierarchical model has the following form:

$$\begin{aligned} \alpha &\sim \pi(\alpha) \\ \theta_1,\ldots,\theta_n|\alpha &\sim p(\theta|\alpha) \\ X_i|\theta_i &\sim p(X_i|\theta_i),\;\; i=1,\ldots n. \end{aligned}$$

As a simple example, suppose that θ_i is the infection rate at hospital i and X_i is presence or absence of infection in a patient at hospital i. It might be reasonable to view the infection rates $\theta_1,\ldots,\theta_n$ as random draws from a distribution $p(\theta\,|\,\alpha)$. This distribution depends on parameters α, known as hyperparameters. We consider hierarchical models in more detail in Example 226.

12.5 Simulation Methods for Bayesian Computation

Suppose that we wish to draw a random sample X from a distribution F. Since $F(X)$ is uniformly distributed over the interval $(0,1)$, a basic strategy is to sample $U \sim$ Uniform$(0,1)$, and then output $X = F^{-1}(U)$. This is an example of simulation; we sample from a distribution that is easy to draw from, in this case Uniform$(0,1)$, and use it to sample from a more complicated distribution F. As another example, suppose that we wish to estimate the integral $\int_0^1 h(\mathrm{x})\, d\mathrm{x}$ for some complicated function h. The basic simulation approach is to draw N samples $X_i \sim$ Uniform$(0,1)$ and estimate the integral as

$$\int_0^1 h(\mathrm{x})\, d\mathrm{x} \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i). \tag{12.8}$$

This converges to the desired integral by the law of large numbers.

Simulation methods are especially useful in Bayesian inference, where complicated distributions and integrals are of the essence; let us briefly review the main ideas. Given a prior $\pi(\theta)$ and data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ the posterior density is

$$\pi(\theta \,|\, \mathcal{D}_n) = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c} \tag{12.9}$$

where $\mathcal{L}_n(\theta)$ is the likelihood function and

$$c = \int \mathcal{L}_n(\theta)\pi(\theta)\, d\theta \tag{12.10}$$

is the normalizing constant. The posterior mean is

$$\overline{\theta} = \int \theta\pi(\theta \,|\, \mathcal{D}_n) d\theta = \frac{1}{c}\int \theta\mathcal{L}_n(\theta)\pi(\theta) d\theta. \tag{12.11}$$

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ is multidimensional, then we might be interested in the posterior for one of the components, θ_1, say. This marginal posterior density is

$$\pi(\theta_1 \,|\, \mathcal{D}_n) = \int\int \cdots \int \pi(\theta_1, \ldots, \theta_k \,|\, \mathcal{D}_n) d\theta_2 \cdots d\theta_k \tag{12.12}$$

which involves high-dimensional integration. When θ is high-dimensional, it may not be feasible to calculate these integrals analytically. Simulation methods will often be helpful.

12.5.1 Basic Monte Carlo Integration

Suppose we want to evaluate the integral

$$I = \int_a^b h(\mathrm{x})\, d\mathrm{x} \tag{12.13}$$

for some function h. If h is an "easy" function like a polynomial or trigonometric function, then we can do the integral in closed form. If h is complicated there may be no known closed form expression for I. There are many numerical techniques for evaluating I such as Simpson's rule, the trapezoidal rule and Gaussian quadrature. Monte Carlo integration is another approach for approximating I which is notable for its simplicity, generality and scalability.

Begin by writing

$$I = \int_a^b h(\mathrm{x}) d\mathrm{x} = \int_a^b w(\mathrm{x}) f(\mathrm{x}) d\mathrm{x} \tag{12.14}$$

where $w(\mathrm{x}) = h(\mathrm{x})(b-a)$ and $f(\mathrm{x}) = 1/(b-a)$. Notice that f is the probability density for a uniform random variable over (a, b). Hence,

$$I = \mathbb{E}_f(w(X)) \tag{12.15}$$

where $X \sim \text{Uniform}(a, b)$. If we generate $X_1, \ldots, X_N \sim \text{Uniform}(a, b)$, then by the law of large numbers

$$\widehat{I} \equiv \frac{1}{N} \sum_{i=1}^N w(X_i) \xrightarrow{P} \mathbb{E}(w(X)) = I. \tag{12.16}$$

This is the basic Monte Carlo integration method. We can also compute the standard error of the estimate

$$\widehat{\mathrm{se}} = \frac{s}{\sqrt{N}} \tag{12.17}$$

where

$$s^2 = \frac{\sum_{i=1}^N (Y_i - \widehat{I})^2}{N-1} \tag{12.18}$$

where $Y_i = w(X_i)$. A $1-\alpha$ confidence interval for I is $\widehat{I} \pm z_{\alpha/2}\widehat{\mathrm{se}}$. We can take N as large as we want and hence make the length of the confidence interval very small.

Example 219. Let $h(\mathrm{x}) = \mathrm{x}^3$. Then, $I = \int_0^1 \mathrm{x}^3 d\mathrm{x} = 1/4$. Based on $N = 10,000$ observations from a Uniform$(0, 1)$ we get $\widehat{I} = .248$ with a standard error of .0028.

A generalization of the basic method is to consider integrals of the form

$$I = \int_a^b h(\mathrm{x}) f(\mathrm{x}) d\mathrm{x} \tag{12.19}$$

where $f(\mathrm{x})$ is a probability density function. Taking f to be a Uniform(a, b) gives us the special case above. Now we draw $X_1, \ldots, X_N \sim f$ and take

$$\widehat{I} := \frac{1}{N} \sum_{i=1}^N h(X_i) \tag{12.20}$$

as before.

Example 220. Let

$$f(\mathrm{x}) = \frac{1}{\sqrt{2\pi}} e^{-\mathrm{x}^2/2} \tag{12.21}$$

be the standard normal PDF. Suppose we want to compute the CDF at some point x:

$$I = \int_{-\infty}^{\mathrm{x}} f(s)ds = \Phi(\mathrm{x}). \tag{12.22}$$

Write

$$I = \int h(s)f(s)ds \tag{12.23}$$

where

$$h(s) = \begin{cases} 1 & s < x \\ 0 & s \geq x. \end{cases} \tag{12.24}$$

Now we generate $X_1, \ldots, X_N \sim N(0, 1)$ and set

$$\widehat{I} = \frac{1}{N} \sum_i h(X_i) = \frac{\text{number of observations } \leq x}{N}. \tag{12.25}$$

For example, with $\mathrm{x} = 2$, the true answer is $\Phi(2) = .9772$ and the Monte Carlo estimate with $N = 10,000$ yields .9751. Using $N = 100,000$ we get .9771.

Example 221 (Bayesian inference for two binomials)**.** Let $X \sim \mathrm{Binomial}(n, p_1)$ and $Y \sim \mathrm{Binomial}(m, p_2)$. We would like to estimate $\delta = p_2 - p_1$. The MLE is $\widehat{\delta} = \widehat{p}_2 - \widehat{p}_1 = (Y/m) - (X/n)$. We can get the standard error $\widehat{\mathrm{se}}$ using the delta method, which yields

$$\widehat{\mathrm{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} \tag{12.26}$$

and then construct a 95 percent confidence interval $\widehat{\delta} \pm 2\,\widehat{\mathrm{se}}$. Now consider a Bayesian analysis. Suppose we use the prior $\pi(p_1, p_2) = \pi(p_1)\pi(p_2) = 1$, that is, a flat prior on (p_1, p_2). The posterior is

$$\pi(p_1, p_2 \,|\, X, Y) \propto p_1^X (1-p_1)^{n-X}\, p_2^Y (1-p_2)^{m-Y}. \tag{12.27}$$

The posterior mean of δ is

$$\overline{\delta} = \int_0^1 \int_0^1 \delta(p_1, p_2)\, \pi(p_1, p_2 \,|\, X, Y) = \int_0^1 \int_0^1 (p_2 - p_1)\, \pi(p_1, p_2 \,|\, X, Y). \tag{12.28}$$

If we want the posterior density of δ we can first get the posterior CDF

$$F(c \,|\, X, Y) = \mathbb{P}(\delta \leq c \,|\, X, Y) = \int_A \pi(p_1, p_2 \,|\, X, Y) \tag{12.29}$$

where $A = \{(p_1, p_2) : \ p_2 - p_1 \le c\}$, and then differentiate F. But this is complicated; to avoid all these integrals, let's use simulation.

Note that $\pi(p_1, p_2 \,|\, X, Y) = \pi(p_1 \,|\, X)\, \pi(p_2 \,|\, Y)$ which implies that p_1 and p_2 are independent under the posterior distribution. Also, we see that $p_1 \,|\, X \sim \mathrm{Beta}(X + 1, n - X + 1)$ and $p_2 \,|\, Y \sim \mathrm{Beta}(Y + 1, m - Y + 1)$. Hence, we can simulate $(P_1^{(1)}, P_2^{(1)}), \ldots, (P_1^{(N)}, P_2^{(N)})$ from the posterior by drawing

$$P_1^{(i)} \sim \mathrm{Beta}(X + 1, n - X + 1) \tag{12.30}$$

$$P_2^{(i)} \sim \mathrm{Beta}(Y + 1, m - Y + 1) \tag{12.31}$$

for $i = 1, \ldots, N$. Now let $\delta^{(i)} = P_2^{(i)} - P_1^{(i)}$. Then,

$$\overline{\delta} \approx \frac{1}{N} \sum_{i=1}^{N} \delta^{(i)}. \tag{12.32}$$

We can also get a 95 percent posterior interval for δ by sorting the simulated values, and finding the $.025$ and $.975$ quantile. The posterior density $f(\delta \,|\, X, Y)$ can be obtained by applying density estimation techniques to $\delta^{(1)}, \ldots, \delta^{(N)}$ or, simply by plotting a histogram. For example, suppose that $n = m = 10$, $X = 8$ and $Y = 6$. From a posterior sample of size 1000 we get a 95 percent posterior interval of $(-0.52, 0.20)$. The posterior density can be estimated from a histogram of the simulated values as shown in Figure 12.5.

Example 222 (Bayesian inference for dose response)**.** Suppose we conduct an experiment by giving rats one of ten possible doses of a drug, denoted by $\mathrm{x}_1 < \mathrm{x}_2 < \ldots < \mathrm{x}_{10}$. For each dose level x_i we use n rats and we observe Y_i, the number that survive. Thus we have ten independent binomials $Y_i \sim \mathrm{Binomial}(n, p_i)$. Suppose we know from biological considerations that higher doses should have higher probability of death; thus, $p_1 \le p_2 \le \cdots \le p_{10}$. We want to estimate the dose at which the animals have a 50 percent chance of dying—this is called the LD50. Formally, $\delta = \mathrm{x}_{j^*}$ where

$$j^* = \min \left\{ j : \ p_j \ge \tfrac{1}{2} \right\}. \tag{12.33}$$

Notice that δ is implicitly just a complicated function of $p_1, \ldots, p_{10}$ so we can write $\delta = g(p_1, \ldots, p_{10})$ for some g. This just means that if we know $(p_1, \ldots, p_{10})$ then we can find δ. The posterior mean of δ is

$$\int \int \cdots \int_A g(p_1, \ldots, p_{10})\, \pi(p_1, \ldots, p_{10} \,|\, Y_1, \ldots, Y_{10})\, dp_1 dp_2 \ldots dp_{10}. \tag{12.34}$$

The integral is over the region

$$A = \{(p_1, \ldots, p_{10}) : \ p_1 \le \cdots \le p_{10}\}. \tag{12.35}$$

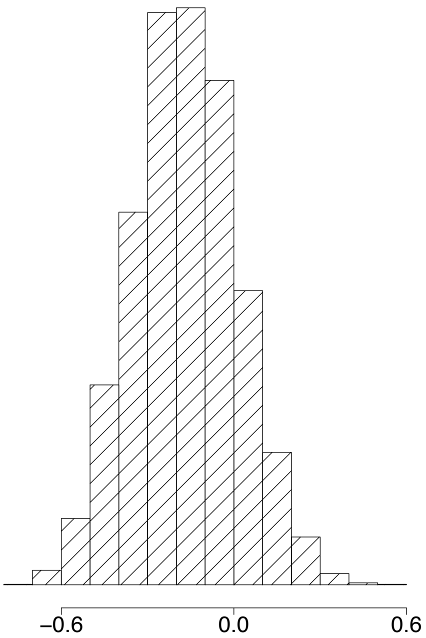


Figure 12.5: Posterior of δ from simulation.

The posterior CDF of δ is

$$\begin{aligned} F(c \,|\, Y_1, \ldots, Y_{10}) &= \mathbb{P}(\delta \le c \,|\, Y_1, \ldots, Y_{10}) && (12.36) \\ &= \int\int \cdots \int_B \pi(p_1, \ldots, p_{10} \,|\, Y_1, \ldots, Y_{10})\, dp_1 dp_2 \ldots dp_{10} && (12.37) \end{aligned}$$

where

$$B = A \cap \Big\{ (p_1, \ldots, p_{10}) : \; g(p_1, \ldots, p_{10}) \le c \Big\}. \qquad (12.38)$$

The posterior mean involves a 10-dimensional integral over a restricted region A. We can approximate this integral using simulation.

Let us take a flat prior truncated over A. Except for the truncation, each P_i has once again a Beta distribution. To draw from the posterior we proceed as follows:

(1) Draw $P_i \sim \mathrm{Beta}(Y_i + 1, n - Y_i + 1), i = 1, \ldots, 10$.

(2) If $P_1 \le P_2 \le \cdots \le P_{10}$ keep this draw. Otherwise, throw it away and draw again until you get one you can keep.

(3) Let $\delta = \mathrm{x}_{j^*}$ where

$$j^* = \min\{j : \; P_j > \tfrac{1}{2}\}. \qquad (12.39)$$

We repeat this N times to get $\delta^{(1)}, \ldots, \delta^{(N)}$ and take

$$\mathbb{E}(\delta \mid Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta^{(i)}. \tag{12.40}$$

Note that δ is a discrete variable. We can estimate its probability mass function by

$$\mathbb{P}(\delta = \mathrm{x}_j \mid Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_{i=1}^{N} I(\delta^{(i)} = \mathrm{x}_j). \tag{12.41}$$

For example, consider the following data:

Dose	1	2	3	4	5	6	7	8	9	10
Number of animals n_i	15	15	15	15	15	15	15	15	15	15
Number of survivors Y_i	0	0	2	2	8	10	12	14	15	14

The posterior draws for $p_1, \ldots, p_{10}$ with $N = 500$ are shown in Figure 12.6. We find that $\overline{\delta} = 5.45$ with a 95 percent interval of (5,7).

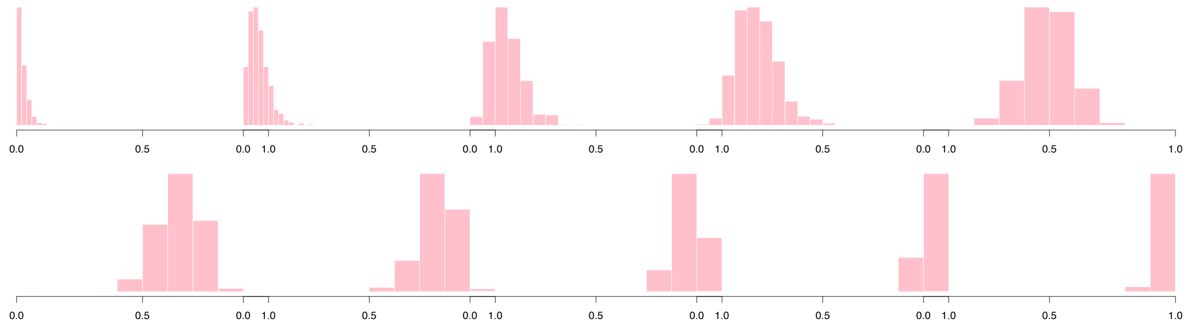


Figure 12.6: Posterior distributions of the probabilities P_i, $i = 1, \ldots, 10$, for the dose response data of Example 222.

12.5.2 Importance Sampling

Consider again the integral $I = \displaystyle\int h(\mathrm{x}) f(\mathrm{x}) d\mathrm{x}$ where f is a probability density. The basic Monte Carlo method involves sampling from f. However, there are cases where we may not know how to sample from f. For example, in Bayesian inference, the posterior density is obtained by multiplying the likelihood $\mathcal{L}_n(\theta)$ times the prior $\pi(\theta)$, and there is generally no guarantee that $\pi(\theta \mid \mathcal{D}_n)$ will be a known distribution like a normal or gamma.

Importance sampling is a generalization of basic Monte Carlo that addresses this problem. Let g be a probability density that we know how to sample from. Then

$$I = \int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)} g(x)dx = \mathbb{E}_g(Y) \tag{12.42}$$

where $Y = h(X)f(X)/g(X)$ and the expectation $\mathbb{E}_g(Y)$ is with respect to g. We can simulate $X_1, \ldots, X_N \sim g$ and estimate I by the sample average

$$\widehat{I} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{N}\sum_{i=1}^{N} \frac{h(X_i)f(X_i)}{g(X_i)}. \tag{12.43}$$

This is called importance sampling. By the law of large numbers, $\widehat{I} \xrightarrow{P} I$.

There's a catch, however. It's possible that $\widehat{I}$ might have an infinite standard error. To see why, recall that I is the mean of $w(\mathrm{x}) = h(\mathrm{x})f(\mathrm{x})/g(\mathrm{x})$. The second moment of this quantity is

$$\mathbb{E}_g(w^2(X)) = \int \left(\frac{h(\mathrm{x})f(\mathrm{x})}{g(\mathrm{x})}\right)^2 g(\mathrm{x})d\mathrm{x} = \int \frac{h^2(\mathrm{x})f^2(\mathrm{x})}{g(\mathrm{x})} d\mathrm{x}. \tag{12.44}$$

If g has thinner tails than f, then this integral might be infinite. To avoid this, a basic rule in importance sampling is to sample from a density g with thicker tails than f. Also, suppose that $g(\mathrm{x})$ is small over some set A where $f(\mathrm{x})$ is large. Again, the ratio of f/g could be large leading to a large variance. This implies that we should choose g to be similar in shape to f. In summary, a good choice for an importance sampling density g should be similar to f but with thicker tails. In fact, we can say what the optimal choice of g is.

Theorem 223. The choice of g that minimizes the variance of $\widehat{I}$ is

$$g^*(\mathrm{x}) = \frac{|h(\mathrm{x})|f(\mathrm{x})}{\int |h(s)|f(s)ds}. \tag{12.45}$$

Proof. The variance of $w = fh/g$ is

$$\begin{aligned}
\mathbb{E}_g(w^2) - (\mathbb{E}(w^2))^2 &= \int w^2(\mathrm{x})g(\mathrm{x})d\mathrm{x} - \left(\int w(\mathrm{x})g(\mathrm{x})d\mathrm{x}\right)^2 &(12.46)\\
&= \int \frac{h^2(\mathrm{x})f^2(\mathrm{x})}{g^2(\mathrm{x})} g(\mathrm{x})d\mathrm{x} - \left(\int \frac{h(\mathrm{x})f(\mathrm{x})}{g(\mathrm{x})} g(\mathrm{x})d\mathrm{x}\right)^2 &(12.47)\\
&= \int \frac{h^2(\mathrm{x})f^2(\mathrm{x})}{g^2(\mathrm{x})} g(\mathrm{x})d\mathrm{x} - \left(\int h(\mathrm{x})f(\mathrm{x})d\mathrm{x}\right)^2. &(12.48)
\end{aligned}$$

The second integral does not depend on g, so we only need to minimize the first integral. From Jensen's inequality, we have

$$\mathbb{E}_g(W^2) \geq (\mathbb{E}_g(|W|))^2 = \left(\int |h(\mathrm{x})| f(\mathrm{x}) d\mathrm{x}\right)^2. \tag{12.49}$$

This establishes a lower bound on $\mathbb{E}_g(W^2)$. However, $\mathbb{E}_{g^*}(W^2)$ equals this lower bound which proves the claim. □

This theorem is interesting but it is only of theoretical interest. If we did not know how to sample from f then it is unlikely that we could sample from $|h(\mathrm{x})|f(\mathrm{x})/\int |h(s)|f(s)ds$. In practice, we simply try to find a thick-tailed distribution g which is similar to $f|h|$.

Example 224 (Tail probability)**.** Let's estimate $I = \mathbb{P}(Z > 3) = .0013$ where $Z \sim N(0,1)$. Write

$$I = \int h(\mathrm{x}) f(\mathrm{x}) d\mathrm{x}$$

where $f(\mathrm{x})$ is the standard normal density and $h(\mathrm{x}) = 1$ if $\mathrm{x} > 3$, and 0 otherwise. The basic Monte Carlo estimator is $\widehat{I} = N^{-1}\sum_i h(X_i)$ where $X_1, \ldots, X_N \sim N(0,1)$. Using $N = 100$ we find (from simulating many times) that $\mathbb{E}(\widehat{I}) = .0015$ and $\mathrm{Var}(\widehat{I}) = .0039$. Notice that most observations are wasted in the sense that most are not near the right tail. Now we will estimate this with importance sampling taking g to be a Normal(4,1) density. We draw values from g and the estimate is now

$$\widehat{I} = N^{-1}\sum_{i=1}^{N} f(X_i)h(X_i)/g(X_i).$$

In this case we find that $\mathbb{E}(\widehat{I}) = .0011$ and $\mathrm{Var}(\widehat{I}) = .0002$. We have reduced the standard deviation by a factor of 20.

To see how importance sampling can be used in Bayesian inference, consider the posterior mean $\overline{\theta} = \mathbb{E}[\theta|X_1, \ldots, X_n]$. Let g be an importance sampling distribution. Then

$$\mathbb{E}[\theta|X_1, \ldots, X_n] = \frac{\int \theta \mathcal{L}(\theta)\pi(\theta)d\theta}{\int \mathcal{L}(\theta)\pi(\theta)d\theta} = \frac{\int h_1(\theta)g(\theta)d\theta}{\int h_2(\theta)g(\theta)d\theta}$$

where

$$h_1(\theta) = \frac{\theta\mathcal{L}(\theta)\pi(\theta)}{g(\theta)}, \quad h_2(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{g(\theta)}.$$

Let $\theta_1, \ldots, \theta_N$ be a sample from g. Then

$$\mathbb{E}[\theta|X_1, \ldots, X_n] \approx \frac{\frac{1}{N}\sum_{i=1}^{N} h_1(\theta_i)}{\frac{1}{N}\sum_{i=1}^{N} h_2(\theta_i)}.$$

This looks very simple but, in practice, it is very difficult to choose a good importance sampler g, especially in high dimensions. With a poor choice of g, the variance of the estimate is huge. This is the main motivation for more modern methods such as MCMC.

Many variants of the basic importance sampling scheme have been proposed and studied; see, for example [65] and [80].

12.5.3 Markov Chain Monte Carlo (MCMC)

Consider once more the problem of estimating the integral $I = \int h(x)f(x)dx$. Now we introduce Markov chain Monte Carlo (MCMC) methods. The idea is to construct a Markov chain $X_1, X_2, \ldots,$ whose stationary distribution is f. Under certain conditions it will then follow that

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) \xrightarrow{P} \mathbb{E}_f(h(X)) = I. \tag{12.50}$$

This works because there is a law of large numbers for Markov chains; see the appendix.

The Metropolis–Hastings algorithm is a specific MCMC method that works as follows. Let $q(\mathrm{y}\,|\,\mathrm{x})$ be an arbitrary, "friendly" distribution—that is, we know how to sample efficiently from $q(\mathrm{y}\,|\,\mathrm{x})$. The conditional density $q(\mathrm{y}\,|\,\mathrm{x})$ is called the proposal distribution. The Metropolis–Hastings algorithm creates a sequence of observations $X_0, X_1, \ldots,$ as follows.

Metropolis–Hastings Algorithm
Choose X_0 arbitrarily.

Given $X_0, X_1, \ldots, X_i$, generate X_{i+1} as follows:

1. Generate a proposal or candidate value $Y \sim q(\mathrm{y}\,|\,X_i)$.
2. Evaluate $r \equiv r(X_i, Y)$ where

$$r(\mathrm{x},\mathrm{y}) = \min\left\{\frac{f(\mathrm{y})}{f(\mathrm{x})}\frac{q(\mathrm{x}\,|\,\mathrm{y})}{q(\mathrm{y}\,|\,\mathrm{x})},\ 1\right\}. \tag{12.51}$$

3. Set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r \\ X_i & \text{with probability } 1-r. \end{cases} \tag{12.52}$$

A simple way to execute step (3) is to generate $U \sim \mathrm{Uniform}(0,1)$. If $U < r$ set $X_{i+1} = Y$; otherwise set $X_{i+1} = X_i$. A common choice for $q(\mathrm{y}\,|\,\mathrm{x})$ is $N(\mathrm{x}, b^2)$ for some $b > 0$, so that the proposal is draw from a normal, centered at the current value. In this case, the proposal density q is symmetric, $q(\mathrm{y}\,|\,\mathrm{x}) = q(\mathrm{x}\,|\,\mathrm{y})$, and r simplifies to

$$r = \min\left\{ \frac{f(Y)}{f(X_i)},\ 1 \right\}. \tag{12.53}$$

By construction, $X_0, X_1, \ldots$ is a Markov chain. But why does this Markov chain have f as its stationary distribution? Before we explain why, let us first do an example.

Example 225. The Cauchy distribution has density

$$f(\mathrm{x}) = \frac{1}{\pi}\frac{1}{1+\mathrm{x}^2}. \tag{12.54}$$

Our goal is to simulate a Markov chain whose stationary distribution is f. As suggested in the remark above, we take $q(\mathrm{y}\,|\,\mathrm{x})$ to be a $N(\mathrm{x}, b^2)$. So in this case,

$$r(\mathrm{x},\mathrm{y}) = \min\left\{ \frac{f(\mathrm{y})}{f(\mathrm{x})},\ 1 \right\} = \min\left\{ \frac{1+\mathrm{x}^2}{1+\mathrm{y}^2},\ 1 \right\}. \tag{12.55}$$

So the algorithm is to draw $Y \sim N(X_i, b^2)$ and set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r(X_i, Y) \\ X_i & \text{with probability } 1 - r(X_i, Y). \end{cases} \tag{12.56}$$

The simulator requires a choice of b. Figure 12.7 shows three chains of length $N = 1,000$ using $b = .1$, $b = 1$ and $b = 10$. Setting $b = .1$ forces the chain to take small steps. As a result, the chain doesn't "explore" much of the sample space. The histogram from the sample does not approximate the true density very well. Setting $b = 10$ causes the proposals to often be far in the tails, making r small and hence we reject the proposal and keep the chain at its current position. The result is that the chain "gets stuck" at the same place quite often. Again, this means that the histogram from the sample does not approximate the true density very well. The middle choice avoids these extremes and results in a Markov chain sample that better represents the density sooner. In summary, there are tuning parameters and the efficiency of the chain depends on these parameters. We'll discuss this in more detail later.

If the sample from the Markov chain starts to look like the target distribution f quickly, then we say that the chain is "mixing well." Constructing a chain that mixes well is somewhat of an art.

12.5.4 Why It Works

An understanding of why MCMC works requires elementary Markov chain theory, which is reviewed in an Appendix at the end of this chapter. We describe a Markov chain with the

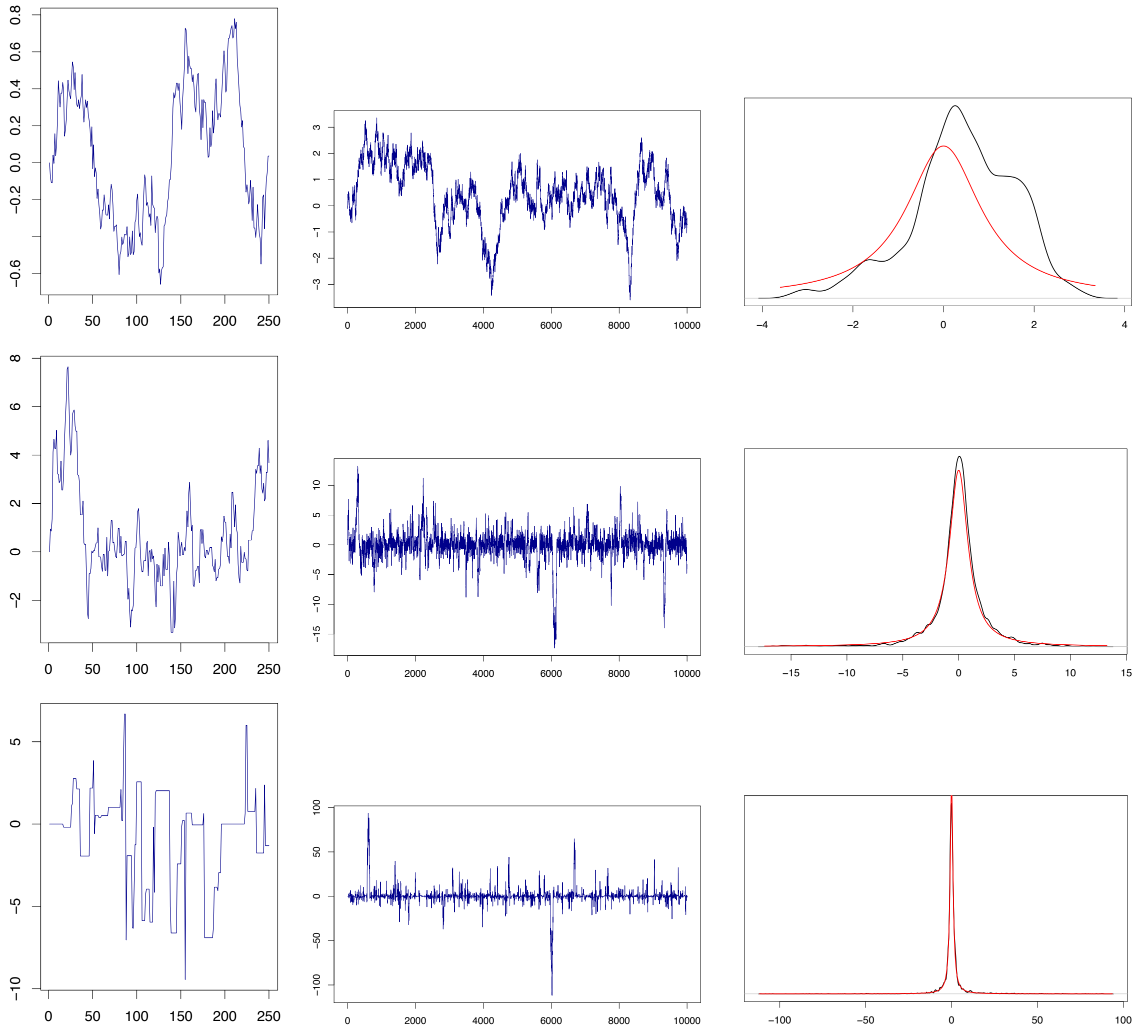


Figure 12.7: Three Metropolis chains corresponding to $b = .1$, $b = 1$, $b = 10$, with acceptance rates 97%, 76%, and 27%, respectively.

transition kernel $p(y|x)$ which is the probability of making a transition from x to y. We say that f is a stationary distribution for the Markov chain if $f(\mathrm{x}) = \int f(\mathrm{y})p(x|y)\, d\mathrm{y}$. This can be interpreted as follows. Once the chain reaches the distribution f it stays in f. Applying another step of the chain $p(x|y)$ does not change the distribution. Under approrpriate conditions, the following is true. If f is a stationary distribution for a Markov chain, then the data from a sample run of the Markov chain will approximate the distribution f. In other words, if we can design a Markov chain with stationary distribution f then we can run the Markov chai and use the resulting data as if it were a sample from f.

We say that the chain satisfies detailed balance holds with respect to f if

$$f(\mathrm{x})p(y|x) = f(\mathrm{y})p(x|y). \tag{12.57}$$

Intuitively, suppose we draw once from f then apply one step of the chain. Deatiled balance means that (x, y) has the same probablity as (y, x). In other words, the chain is time reversible.

If the chian satisfies detailed balance holds with respect to f then f must be a stationary distribution. To see this, note that

$$\int f(\mathrm{y})p(x|y)\,d\mathrm{y} = \int f(\mathrm{x})p(y|x)\,d\mathrm{y} = f(\mathrm{x})\int p(y|x)\,d\mathrm{y} = f(\mathrm{x}) \tag{12.58}$$

which shows that $f(x) = \int f(\mathrm{y})p(x|y)\,d\mathrm{y}$ as required.

Our goal is to show that when $p(y|x)$ is the Markov chain defined by the Metropolis-Hastings algorithm, then f satisfies detailed balance, and therefore is a stationary distribution for the chain.

Recall the the Metropolis-Hastings algorithm uses a user-chosem distribution $q(y|x)$ together with a accept/reject step. Tis defines a Markov chain with transition probablity $p(y|x) = q(y|x)r(x, y)$. Consider two points x and y. Either

$$f(\mathrm{x})q(\mathrm{y}\,|\,\mathrm{x}) < f(\mathrm{y})q(\mathrm{x}\,|\,\mathrm{y}) \quad \text{or} \quad f(\mathrm{x})q(\mathrm{y}\,|\,\mathrm{x}) > f(\mathrm{y})q(\mathrm{x}\,|\,\mathrm{y}). \tag{12.59}$$

We will ignore ties (which occur with probability zero for continuous distributions). Without loss of generality, assume that $f(\mathrm{x})q(\mathrm{y}\,|\,\mathrm{x}) > f(\mathrm{y})q(\mathrm{x}\,|\,\mathrm{y})$. This implies that

$$r(\mathrm{x}, \mathrm{y}) = \frac{f(\mathrm{y})}{f(\mathrm{x})}\frac{q(\mathrm{x}\,|\,\mathrm{y})}{q(\mathrm{y}\,|\,\mathrm{x})} < 1 \tag{12.60}$$

and that $r(\mathrm{y}, \mathrm{x}) = 1$.

Now let $p(y|x)$ be the probability of jumping from x to y. This means that (i) the proposal distribution must generate y, and (ii) you must accept y. Thus,

$$p(y|x) = q(\mathrm{y}\,|\,\mathrm{x})r(\mathrm{x}, \mathrm{y}) = q(\mathrm{y}\,|\,\mathrm{x})\frac{f(\mathrm{y})}{f(\mathrm{x})}\frac{q(\mathrm{x}\,|\,\mathrm{y})}{q(\mathrm{y}\,|\,\mathrm{x})} = \frac{f(\mathrm{y})}{f(\mathrm{x})}q(\mathrm{x}\,|\,\mathrm{y}). \tag{12.61}$$

Therefore,

$$f(\mathrm{x})p(y|x) = f(\mathrm{y})q(\mathrm{x}\,|\,\mathrm{y}). \tag{12.62}$$

On the other hand, $p(x|y)$ is the probability of jumping from y to x. This requires two that (i) the proposal distribution must generate x, and (ii) you must accept x. This occurs with probability $p(\mathrm{y}, \mathrm{x}) = q(\mathrm{x}\,|\,\mathrm{y})r(\mathrm{y}, \mathrm{x}) = q(\mathrm{x}\,|\,\mathrm{y})$. Hence,

$$f(\mathrm{y})p(x|y) = f(\mathrm{y})q(\mathrm{x}\,|\,\mathrm{y}). \tag{12.63}$$

Comparing (12.62) and (12.63), we see that we have shown that detailed balance holds.

12.5.5 Different Flavors of MCMC

There are different types of MCMC algorithm. Here we will consider a few of the most popular versions.

Random-Walk-Metropolis–Hastings. In the previous section we considered drawing a proposal Y of the form

$$Y = X_i + \epsilon_i \tag{12.64}$$

where ϵ_i comes from some distribution with density g. In other words, $q(\mathrm{y}\,|\,\mathrm{x}) = g(\mathrm{y} - \mathrm{x})$. We saw that in this case,

$$r(\mathrm{x}, \mathrm{y}) = \min\left\{1, \frac{f(\mathrm{y})}{f(\mathrm{x})}\right\}. \tag{12.65}$$

This is called a random-walk-Metropolis–Hastings method. The reason for the name is that, if we did not do the accept–reject step, we would be simulating a random walk. The most common choice for g is a $N(0, b^2)$. The hard part is choosing b so that the chain mixes well. As mentioned earlier, a good rule of thumb is to choose b so that about 50 percent of the proposals are accepted.

Note that this method doesn't make sense unless X takes values on the whole real line. If X is restricted to some interval then it is best to transform X. For example, if $X \in (0, \infty)$ then you might take $Y = \log X$ and then simulate the distribution for Y instead of X.

Independence-Metropolis–Hastings. This is an importance-sampling version of MCMC. We draw the proposal from a fixed distribution g. Generally, g is chosen to be an approximation to f. The acceptance probability becomes

$$r(\mathrm{x}, \mathrm{y}) = \min\left\{1, \frac{f(\mathrm{y})}{f(\mathrm{x})}\frac{g(\mathrm{x})}{f(\mathrm{y})}\right\} = \min\left\{1, \frac{f(\mathrm{y})}{g(\mathrm{y})}\frac{g(\mathrm{x})}{f(\mathrm{x})}\right\}. \tag{12.66}$$

Gibbs Sampling. The two previous methods can be easily adapted, in principle, to work in higher dimensions. In practice, tuning the chains to make them mix well is hard. Gibbs sampling is a way to turn a high-dimensional problem into several one-dimensional problems.

Here's how it works for a bivariate problem. Suppose that (X, Y) has density $f_{X,Y}(\mathrm{x}, \mathrm{y})$. First, suppose that it is possible to simulate from the conditional distributions $f_{X\,|\,Y}(\mathrm{x}\,|\,\mathrm{y})$ and $f_{Y\,|\,X}(\mathrm{y}\,|\,\mathrm{x})$. Let (X_0, Y_0) be starting values, and assume we have drawn $(X_0, Y_0), \ldots, (X_n, Y_n)$. Then the Gibbs sampling algorithm for getting (X_{n+1}, Y_{n+1}) is:

To see that this is a special case of the Metropolis-Hastings algorithm, suppose that the current state is (X_n, Y_n) and the proposal is (X_n, Y), with probability $f_{Y\,|\,X}(Y\,|\,X_n)$. Then the acceptance probability in the Metropolis-Hastings algorithm is

$$r((X_n, Y_n), (X_n, Y)) \;=\; \min\left\{1, \frac{f(X_n, Y)}{f(X_n, Y_n)}\frac{f_{Y\,|\,X}(Y_n\,|\,X_n)}{f_{Y\,|\,X}(Y\,|\,X_n)}\right\} \tag{12.69}$$

Gibbs Sampling Iterate until convergence:

$$\begin{aligned} X_{n+1} &\sim f_{X\,|\,Y}(x\,|\,Y_n) && (12.67)\\ Y_{n+1} &\sim f_{Y\,|\,X}(y\,|\,X_{n+1}) && (12.68)\end{aligned}$$

$$= \min\left\{1, \frac{f(X_n,Y)}{f(X_n,Y_n)}\frac{f(X_n,Y_n)}{f(X_n,Y)}\right\} = 1. \qquad (12.70)$$

This generalizes in the obvious way to higher dimensions, where we cycle through the variables, sampling one of them at a time, conditioned on the others.

Example 226 (Normal hierarchical model)**.** Gibbs sampling is very useful for hierarchical models. Here is a simple case. Suppose we have a sample of data from k cities. From each city we draw n_i people and observe how many people Y_i have a disease. Thus, $Y_i \sim \text{Binomial}(n_i, p_i)$, allowing for different disease rates in different cities. We can also think of the p_i's as random draws from some distribution F. We can write this model in the following way:

$$\begin{aligned} P_i &\sim F && (12.71)\\ Y_i\,|\,P_i = p_i &\sim \text{Binomial}(n_i, p_i). && (12.72)\end{aligned}$$

We are interested in estimating the p_i's and the overall disease rate $\int p\,d\pi(p)$.

To proceed, it will simplify matters if we make some transformations that allow us to use some normal approximations. Let $\widehat{p}_i = Y_i/n_i$. Recall that $\widehat{p}_i \approx N(p_i, s_i)$ where $s_i = \sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$. Let $\psi_i = \log(p_i/(1-p_i))$ and define $Z_i \equiv \widehat{\psi}_i = \log(\widehat{p}_i/(1-\widehat{p}_i))$. By the delta method,

$$\widehat{\psi}_i \approx N(\psi_i, \sigma_i^2) \qquad (12.73)$$

where $\sigma_i^2 = 1/(n\widehat{p}_i(1-\widehat{p}_i))$. Experience shows that the normal approximation for ψ is more accurate than the normal approximation for p so we shall work with ψ, treating σ_i as known. Furthermore, we shall take the distribution of the ψ_i's to be normal. The hierarchical model is now

$$\psi_i \sim N(\mu, \tau^2) \text{ and } Z_i\,|\,\psi_i \sim N(\psi_i, \sigma_i^2). \qquad (12.74)$$

As yet another simplification we take $\tau = 1$. The unknown parameters are $\theta = (\mu, \psi_1, \ldots, \psi_k)$. The likelihood function is

$$\begin{aligned} \mathcal{L}_n(\theta) &\propto \prod_i f(\psi_i\,|\,\mu)\prod_i f(Z_i\,|\,\psi_i) && (12.75)\\ &\propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\sigma_i^2}(Z_i-\psi_i)^2\right\}. && (12.76)\end{aligned}$$

If we use the prior $f(\mu) \propto 1$ then the posterior is proportional to the likelihood. To use Gibbs sampling, we need to find the conditional distribution of each parameter conditional on all the others. Let us begin by finding $f(\mu \,|\, \text{rest})$ where "rest" refers to all the other variables. We can throw away any terms that don't involve μ. Thus,

$$\begin{aligned} f(\mu \,|\, \text{rest}) &\propto \prod_i \exp\left\{ -\frac{1}{2}(\psi_i - \mu)^2 \right\} & (12.77) \\ &\propto \exp\left\{ -\frac{k}{2}(\mu - b)^2 \right\} & (12.78) \end{aligned}$$

where

$$b = \frac{1}{k}\sum_i \psi_i. \tag{12.79}$$

Hence we see that $\mu \,|\, \text{rest} \sim N(b, 1/k)$. Next we will find $f(\psi \,|\, \text{rest})$. Again, we can throw away any terms not involving ψ_i, leaving us with

$$\begin{aligned} f(\psi_i \,|\, \text{rest}) &\propto \exp\left\{ -\frac{1}{2}(\psi_i - \mu)^2 \right\} \exp\left\{ -\frac{1}{2\sigma_i^2}(Z_i - \psi_i)^2 \right\} & (12.80) \\ &\propto \exp\left\{ -\frac{1}{2d_i^2}(\psi_i - e_i)^2 \right\} & (12.81) \end{aligned}$$

where

$$e_i = \frac{\frac{Z_i}{\sigma_i^2} + \mu}{1 + \frac{1}{\sigma_i^2}} \quad \text{and} \quad d_i^2 = \frac{1}{1 + \frac{1}{\sigma_i^2}} \tag{12.82}$$

and so $\psi_i \,|\, \text{rest} \sim N(e_i, d_i^2)$. The Gibbs sampling algorithm then involves iterating the following steps N times:

$$\begin{aligned} \text{draw } \mu &\sim N(b, v^2) & (12.83) \\ \text{draw } \psi_1 &\sim N(e_1, d_1^2) & (12.84) \\ \vdots \quad & \quad \vdots & (12.85) \\ \text{draw } \psi_k &\sim N(e_k, d_k^2). & (12.86) \end{aligned}$$

It is understood that at each step, the most recently drawn version of each variable is used.

We generated a numerical example with $k = 20$ cities and $n = 20$ people from each city. After running the chain, we can convert each ψ_i back into p_i by way of $p_i = e^{\psi_i}/(1 + e^{\psi_i})$. The raw proportions are shown in Figure 12.9. Figure 12.8 shows "trace plots" of the Markov chain for p_1 and μ. Figure 12.9 shows the posterior for μ based on the simulated values. The second panel of Figure 12.9 shows the raw proportions and the Bayes estimates. Note that the Bayes estimates are "shrunk" together. The parameter τ controls the amount of shrinkage. We set $\tau = 1$ but, in practice, we should treat τ as another unknown parameter and let the data determine how much shrinkage is needed.

So far we assumed that we know how to draw samples from the conditionals $f_{X\,|\,Y}(\mathrm{x}\,|\,\mathrm{y})$ and $f_{Y\,|\,X}(\mathrm{y}\,|\,\mathrm{x})$. If we don't know how, we can still use the Gibbs sampling algorithm by drawing each observation using a Metropolis–Hastings step. Let q be a proposal distribution for x and let $\widetilde{q}$ be a proposal distribution for y. When we do a Metropolis step for X, we treat Y as fixed. Similarly, when we do a Metropolis step for Y, we treat X as fixed. Here are the steps:

Gibbs sampling with Metropoplis-Hastings

(1a) Draw a proposal $Z \sim q(z\,|\,X_n)$.

(1b) Evaluate

$$r = \min\left\{ \frac{f(Z,Y_n)}{f(X_n,Y_n)} \frac{q(X_n\,|\,Z)}{q(Z\,|\,X_n)},\ 1 \right\}. \tag{12.87}$$

(1c) Set

$$X_{n+1} = \begin{cases} Z & \text{with probability } r \\ X_n & \text{with probability } 1-r. \end{cases} \tag{12.88}$$

(2a) Draw a proposal $Z \sim \widetilde{q}(z\,|\,Y_n)$.

(2b) Evaluate

$$r = \min\left\{ \frac{f(X_{n+1},Z)}{f(X_{n+1},Y_n)} \frac{\widetilde{q}(Y_n\,|\,Z)}{\widetilde{q}(Z\,|\,Y_n)},\ 1 \right\}. \tag{12.89}$$

(2c) Set

$$Y_{n+1} = \begin{cases} Z & \text{with probability } r \\ Y_n & \text{with probability } 1-r. \end{cases} \tag{12.90}$$

Note that in step (1) (and similarly for step (2)), with Y_n fixed, sampling from $f(Z\,|\,Y_n)$ is equivalent to sampling from $f(Z,Y_n)$, as the ratios are identical:

$$\frac{f(Z,Y_n)}{f(X_n,Y_n)} = \frac{f(Z\,|\,Y_n)}{f(X_n\,|\,Y_n)}. \tag{12.91}$$

12.5.6 Normalizing Constants

The beauty of MCMC is that we avoid having to compute the normalizing constant

$$c = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta.$$

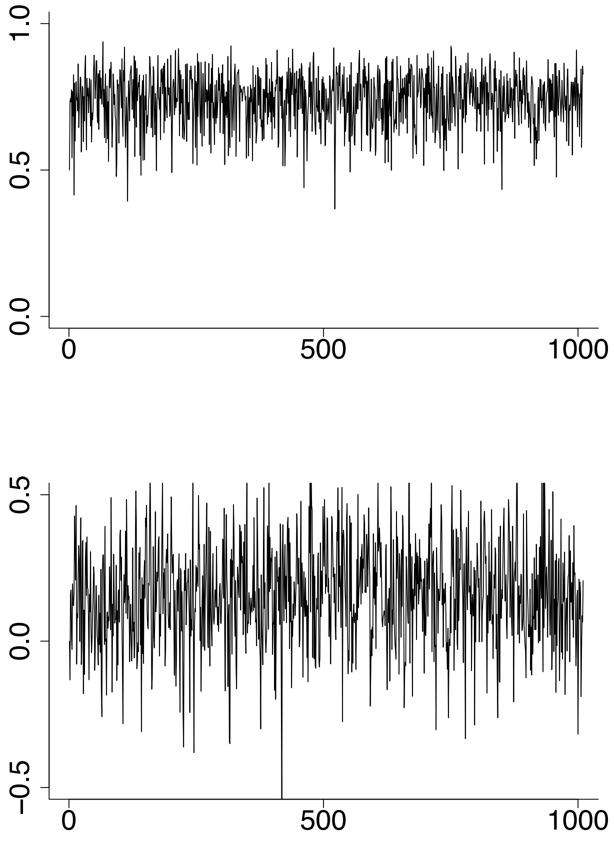


Figure 12.8: Posterior simulation for Example 226. The top panel shows simulated values of p_1. The bottom panel shows simulated values of μ.

But suppose we do want to estimate c. For example, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are two models then

$$\mathbb{P}(\mathcal{M}_1 \,|\, X_1, \ldots, X_n) = \frac{c_1 p}{c_1 p + c_2(1-p)} \tag{12.92}$$

where p is the prior probability of model 1 and c_1, c_2 are the normalizing constants for the two models. Thus, to do Bayesian model selection requires the normalizing constants.

In general, suppose that f is a probability density function and that

$$f(\theta) = cg(\theta) \tag{12.93}$$

where $g(\theta) > 0$ is a known function and c is unknown; typically, $g(\theta) = \mathcal{L}_n(\theta)\pi(\theta)$. Let $\theta_1, \ldots, \theta_n$ be a sample from f. Let h be a known probability density function. Define

$$\widehat{c} = \frac{1}{n}\sum_{i=1}^{n} \frac{h(\theta_i)}{g(\theta_i)}. \tag{12.94}$$

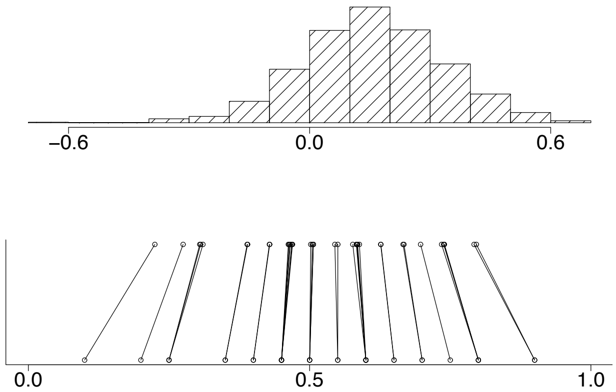


Figure 12.9: Example 226. Top panel: posterior histogram of μ. Lower panel: raw proportions and the Bayes posterior estimates. The Bayes estimates have been shrunk closer together than the raw proportions.

Then

$$\mathbb{E}(\widehat{c}) = \int \frac{h(\theta)}{g(\theta)} f(\theta) d\theta = \int \frac{h(\theta)}{g(\theta)} c g(\theta) d\theta = c. \tag{12.95}$$

And if $\int h^2(\theta)/g(\theta) d\theta < \infty$, then $\widehat{c} - c = O_P(n^{-1/2})$.

12.6 Examples Where Bayesian Inference and Frequentist Inference Disagree

Bayesian inference is appealing when prior information is available since Bayes' theorem is a natural way to combine prior information with data. However, Bayesian inference is also controversial because it inherently embraces a subjective notion of probability. In general, Bayesian methods provide no guarantees on long run performance. In some cases, Bayesian methods can have poor frequency behavior.

Example 227. Normal means. Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ be the data obtained from the model $X \sim N(\mu_i, 1)$. Suppose we use the flat prior $\pi(\mu_1, \ldots, \mu_n) \propto 1$. Then, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, the posterior for $\boldsymbol{\mu}$ is multivariate Normal with mean $\mathbb{E}(\boldsymbol{\mu} \,|\, \mathcal{D}_n) = (X_1, \ldots, X_n)$ and covariance equal to the identity matrix. Let $\theta = \sum_{i=1}^n \mu_i^2$. Let $C_n = [c_n, \infty)$ where c_n is chosen so that $\mathbb{P}(\theta \in C_n \,|\, \mathcal{D}_n) = 0.95$. How often, in the frequentist sense, does C_n trap θ? Stein (1959) showed that

$$\mathbb{P}_\mu(\theta \in C_n) \to 0, \quad \text{as } n \to \infty.$$

Thus, there is a sharp difference between $\mathbb{P}_\mu(\theta \in C_n)$ and $\mathbb{P}(\theta \in C_n \,|\, \mathcal{D}_n)$.

Example 228. Sampling to a Foregone Conclusion. Let $X \sim N(\theta, 1)$ be a univariate random variable and $\mathcal{D}_N = \{X_1, \ldots, X_N\}$ be the observed data. Think of X_i as some statistics that compares a new drug to a placebo. Suppose we continue sampling until $T_N > k$ where $T_N = \sqrt{N}\overline{X}_N$ and k is a fixed number, say, $k = 10$. This means that we stop when the drug appears to be much better than the placebo.

The sample size N is now a random variable. It can be shown that $\mathbb{P}(N < \infty) = 1$. It can also be shown that the posterior $p(\theta \,|\, X_1, \ldots, X_N)$ is the same as if N had been fixed in advance. That is, the randomness in N does not affect the posterior. Now if the prior $\pi(\theta)$ is smooth then the posterior is approximately $\theta \,|\, X_1, \ldots, X_N \sim N(\overline{X}_N, 1/N)$. Hence, if $C_N = \overline{X}_N \pm 1.96/\sqrt{N}$ then $\mathbb{P}(\theta \in C_N \,|\, X_1, \ldots, X_N) \approx 0.95$. Notice that 0 is never in C_N since, when we stop sampling, $T > 10$, and therefore

$$\overline{X}_N - \frac{1.96}{\sqrt{N}} > \frac{10}{\sqrt{N}} - \frac{1.96}{\sqrt{N}} > 0. \tag{12.96}$$

Hence, when $\theta = 0$, $\mathbb{P}_\theta(\theta \in C_N) = 0$. Thus, the frequentist coverage is

$$\text{Coverage} = \inf_\theta \mathbb{P}_\theta(\theta \in C_N) = 0.$$

This is called sampling to a foregone conclusion and is a serious issue in sequential clinical trials.

Example 229 (Godambe's Example)**.** This example is due to V.P. Godambe. Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a finite set of constants. For simplicity, assume that $c_j \in \{0, 1\}$ (although this is not important). Let $\theta = N^{-1}\sum_{j=1}^N c_j$. Suppose we want to estimate θ. We proceed as follows. Let $S_1, \ldots, S_n \sim \text{Bernoulli}(\pi)$ where π is known. If $S_i = 1$ you get to see c_i. Otherwise, you do not. (This is an example of survey sampling.) The likelihood function is

$$\prod_i \pi^{S_i}(1 - \pi)^{1 - S_i}.$$

The unknown parameter does not appear in the likelihood. In fact, there are no unknown parameters in the likelihood. The likelihood function contains no information at all. Hence, the posterior for θ is equal to the prior for θ. No learning occurs from the data.

We can estimate θ easily from a frequentist perspective. Let

$$\widehat{\theta} = \frac{1}{N\pi}\sum_{j=1}^N c_j S_j.$$

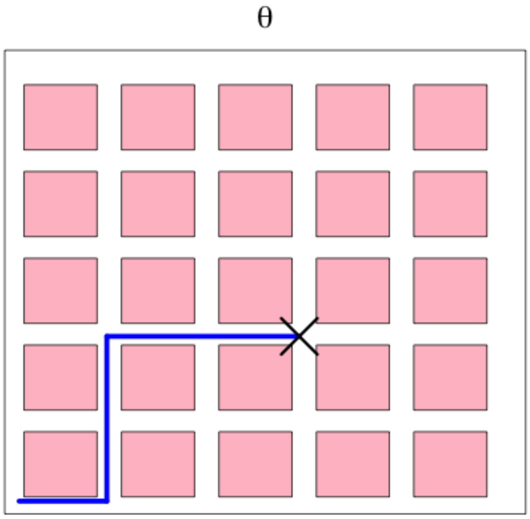
Then $\mathbb{E}(\widehat{\theta}) = \theta$. Hoeffding's inequality implies that

$$\mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) \leq 2e^{-2n\epsilon^2\pi^2}.$$

Hence, $\widehat{\theta}$ is close to θ with high probability.

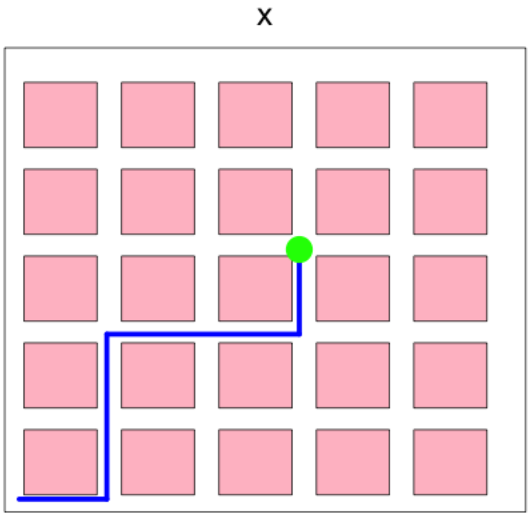
Example 230 (Flatland (Stone's Paradox))**.** Mervyn Stone is Emeritus Professor at University College London. He is famous for his deep work on Bayesian inference as well as pioneering work on cross-validation, coordinate-free multivariate analysis, as well as many other topics. He has a famous example described in Stone (1970, 1976, 1982). In technical jargon, he shows that "a finitely additive measure on the free group with two generators is nonconglomerable." In English: even for a simple problem with a discrete parameters space, Bayesian inference can lead to surprises.

Suppose that Alice wonders randomly in a two dimensional grid-world. She drags an elastic string with her. The string is taut: if she backs up, the string leaves no slack. She can only move in four directions: North, South, West, East. She wandera around for awhile then she stops and buries a treasure. Call this path θ. Here is an example:



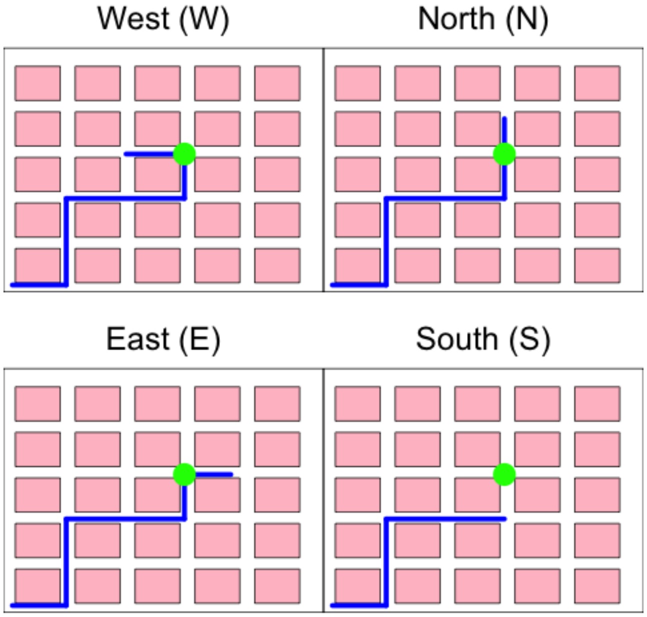
My Path. The X is the treasure.

Now Alice takes one more random step. Each direction has equal probability. Call this path x. So it might look like this:



One more random step.

Two people, Bob (a Bayesian) and Carla (a classical statistician) want to find the treasure. There are only four possible paths that could have yielded x, namely:



Let us call these four paths N, S, W, E. The likelihood is the same for each of these. That is, $p(x|\theta) = 1/4$ for $\theta \in \{N, S, W, E\}$. Suppose Bob uses a flat prior. Since the likelihood is also flat, his posterior is

$$P(\theta = N|x) = P(\theta = S|x) = P(\theta = W|x) = P(\theta = E|x) = \frac{1}{4}.$$

Let B be the three paths that extend x. In this example, $B = \{N, W, E\}$. Then $P(\theta \in B|x) = 3/4$.

Now Carla is very confident and selects a confidence set with only one path, namely, the path that shortens x. In other words, Carla's confidence set is $C = B^c$.

Notice the following strange thing: no matter what θ is, Carla gets the treasure with probability 3/4 while Bob gets the treasure with probability 1/4. That is, $P(\theta \in B|x) = 3/4$ but the coverage of B is 1/4. The coverage of C is 3/4.

Here is quote from Stone (1976): (except that we changed his B and C to Bob and Carla):

" ... it is clear that when Bob and Carla repeatedly engage in this treasure hunt, Bob will find that his posterior probability assignment becomes increasingly discrepant with his proportion of wins and that Carla is, somehow, doing better than [s]he ought. However, there is no message ... that will allow Bob to escape from his Promethean situation; he cannot learn from his experience because each hunt is independent of the other."

Let A be the event that the final step reduced the length of the string. Using the posterior above, we see that Bob finds that $P(A|x) = 3/4$ for each x. Since this holds for each x, Bob deduces that $P(A) = 3/4$. On the other hand, Bob notes that $P(A|\theta) = 1/4$ for every θ. Hence, $P(A) = 1/4$. Bob has just proved that $3/4 = 1/4$.

The apparent contradiction stems from the fact that the prior is improper. Technically this is an example of the non-conglomerability of finitely additive measures. For a rigorous explanation of why this happens you should read Stone's papers. Here is an abbreviated explanation, from Kass and Wasserman (1996, Section 4.2.1).

Let π denotes Bob's improper flat prior and let $\pi(\theta|x)$ denote his posterior distribution. Let π_p denote the prior that is uniform on the set of all paths of length p. This is of course a proper prior. For any fixed x, $\pi_p(A|x) \to 3/4$ as $p \to \infty$. So Bob can claim that his posterior distribution is a limit of well-defined posterior distributions. However, we need to look at this more closely. Let $m_p(x) = \sum_\theta f(x|\theta)\pi_p(\theta)$ be the marginal of x induced by π_p. Let X_p denote all x's of length p or $p + 1$. When $x \in X_p$, $\pi_p(\theta|x)$ is a poor approximation to $\pi(\theta|x)$ since the former is concentrated on a single point while the latter is concentrated on four points. In fact, the total variation distance between $\pi_p(\theta|x)$ and $\pi(\theta|x)$ is 3/4 for $x \in X_p$. (Recall that the total variation distance between two probability measures P and Q is $d(P, Q) = \sup_A |P(A) - Q(A)|$.) Furthermore, X_p is a set with high probability: $m_p(X_p) \to 2/3$ as $p \to \infty$.

While $\pi_p(\theta|x)$ converges to $\pi(\theta|x)$ as $p \to \infty$ for any fixed x, they are not close with high probability.

Here is another description of the problem. Consider a four sided die whose sides are labeled with the symbols $\{a, b, a^{-1}, b^{-1}\}$. We roll the die several times and we record the label on the lowermost face (there is a no uppermost face on a four-sided die). A typical outcome might look like this string of symbols:

$$a \;\; a\, b\, a^{-1}\, b\, b^{-1}\, b\, a\, a^{-1}\, b$$

Now we apply an annihilation rule. If a and a^{-1} appear next to each other, we eliminate these two symbols. Similarly, if b and b^{-1} appear next to each other, we eliminate those two symbols. So the sequence above gets reduced to:

$$a \;\; a\, b\, a^{-1}\, b\, b$$

Let us denote the resulting string of symbols, after removing annihilations, by θ. Now we toss the die one more time. We add this last symbol to θ and we apply the annihilation rule once more. This results in a string which we will denote by x.

You get to see x and you want to infer θ.

Having observed x, there are four possible values of θ and each has the same likelihood. For example, suppose $x = (a, a)$. Then θ has to be one of the following:

$$(a), \;\; (a\,a\,a), \;\; (a\,a\,b^{-1}), \;\; (a\,a\,b)$$

The likelihood function is constant over these four values.

Suppose we use a flat prior on θ. Then the posterior is uniform on these four possibilities. Let $C = C(x)$ denote the three values of θ that are longer than x. Then the posterior satisfies

$$P(\theta \in C|x) = 3/4.$$

Thus $C(x)$ is a 75 percent posterior confidence set.

However, the frequentist coverage of $C(x)$ is 1/4. To see this, fix any θ. Now note that $C(x)$ contains θ if and only if θ concatenated with x is smaller than θ. This happens only if the last symbol is annihilated, which occurs with probability 1/4.

So far we have used a flat prior on the set of paths. One can put a proper prior on the set of paths and compute the posterior but, as we now explain, this does not really help. First of all, if the prior is proper but very flat, you will get a posterior very similar to the posterior from the uniform prior and not much changes. On the other hand, it is possible to choose a specially deisgned prior so that the posterior mimics the freqentist answer. But this poses a problem. If one chooses a prior to represent one's beliefs then it will not give the good, winning behavior of the frequentist method. But if one choose a prior specifically to get a posterior that approximates the frequentist answer, then there is no point of doing Bayesian inference. You might as well just use the frequentist method.

Stone, M. (1970). Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *The Annals of Mathematical Statistics*, 41, 1349-1353,

Stone, M. (1976). Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71, 114-116.

Stone, M. (1982). Review and analysis of some inconsistencies related to improper priors and finite additivity. *Studies in Logic and the Foundations of Mathematics*, 104, 413-426.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343-1370.

12.7 Freedman's Theorem

Here we discuss and interesting result by David Freedman (Annals of Mathematical Statistics, Volume 36, Number 2 (1965), 454-456). The result gets very little attention. Most researchers in statistics and machine learning seem to be unaware of the result. The result says that, "almost all" Bayesian posterior distributions are inconsistent, in a sense we'll make precise below. The math is uncontroversial but, as you might imagine, the intepretation of the result is likely to be controversial.

Let $X_1, \ldots, X_n$ be an iid sample from a distribution P on the natural numbers $I = \{1, 2, 3, \ldots, \}$. Let $\mathcal{P}$ be the set of all such distributions. We endow $\mathcal{P}$ with the weak^* topology. This topology can be described as follows: we say that $P_n \to P$ in the weak^* topology iff $P_n(i) \to P(i)$ for all i.

Let μ denote a prior distribution on $\mathcal{P}$. (More precisely, a prior on an appropriate σ-field.) Let Π be all priors. Again, we endow the set with the weak^* topology. This means that $\mu_n \to \mu$ iff $\int f d\mu_n \to \int f d\mu$ for all bounded, continuous, real functions f.

Let μ_n be the posterior corresponding to the prior μ after n observations. We will say that the pair (P, μ) is consistent if

$$P^\infty(\lim_{n\to\infty} \mu_n = \delta_P) = 1$$

where P^∞ is the product measure corresponding to P and δ_P is a point mass at P.

Now we need to recall some topology. A set is nowhere dense if its closure has an empty interior. A set is meager (or first category) if it is a countable union of nowehere dense sets. Meager sets are small; think of a meager set as the topological version of a null set in measure theory.

Theorem 231 (Freedman 1965)**.** The sets of consistent pairs (P, μ) is meager.

This means that, in a topological sense, consistency is rare for Bayesian procedures. From this result, it can also be shown that most pairs of priors lead to inferences that disagree. (The agreeing pairs are meager.) Or as Freedman says in his paper: " ... it is easy to prove that for essentially any pair of Bayesians, each thinks the other is crazy."

Now, it is possible to choose a prior that will guarantee consistency in the frequentist sense. However, Freedman's theorem says that such priors are rare. Why would a Bayesian choose such a prior? If they choose the prior just to get consistency, this suggests that they are realy trying to be frequentists. If they choose a prior that truly represents their beliefs, then Freedman's theorem implies that the posterior will likely be inconsistent.

12.8 The Bayes-Frequentist Debate

What should we conclude from all this? The important thing is to understand that frequentist and Bayesian methods are answering different questions. Much of the debate about Bayesian and frequentist inference stems from the fact that people confuse the two. We can summarize this as follow:

To analyze subjective beliefs in a principled way: use Bayesian inference.

To design methods with long run frequence guarantees: ue frequentist inference.

In general, Bayesian methods does not have good freqency performance and freqentist methods to do represent anyone's subjective beliefs. They are different tools and there is no reason they should be the same. As it happenns, in low-dimensional models with lots of data, they do tend to be similar. But in high-dimensioal models they are quite different. Generally, Bayesian methods have poor frequentist behavior when the parameter space is high dimensional.

12.9 Summary

The following table presents some comparisons between the two approaches.

	Bayesian	Frequentist
Probability	degree of subjective belief	limiting relative frequency
Typical point estimate	posterior mean or mode	(penalized) maximum likelihood
Optimal estimate	Bayes rule	minimax estimate
Interval estimate	credible interval	confidence interval

12.10 Bibliographic Remarks

Some references on Bayesian inference include [13], [34], [54], [72], and [75]. See [18], [23], [31], [6], [35], [78], and [99] for discussions of some of the technicalities of nonparametric Bayesian inference. See [8] and [46] for a discussion of Bayesian testing. See [47] for a discussion of noninformative priors.