# Some of Bayesian Statistics: The Essential Parts

Rebecca C. Steorts

February 21, 2016

# Contents

# Chapter 1

# Introduction

*There are three kinds of lies: lies, damned lies and statistics.*
—Mark Twain

The word "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace was among the first thinkers to consider the laws of chance and randomness in a quantitative, scientific way. Both Bayes and Laplace were aware of a relation that is now known as Bayes Theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1.1}$$

The proportionality $\propto$ in Eq. (1.1) signifies that the $1/p(x)$ factor is constant and may be ignored when viewing $p(\theta|x)$ as a function of $\theta$. We can decompose Bayes' Theorem into three principal terms:

| | |
|---|---|
| $p(\theta\|x)$ | posterior |
| $p(x\|\theta)$ | likelihood |
| $p(\theta)$ | prior |

In effect, Bayes' Theorem provides a general recipe for updating prior beliefs about an unknown parameter $\theta$ based on observing some data $x$.

However, the notion of having prior beliefs about a parameter that is ostensibly "unknown" did not sit well with many people who considered the problem in the 19th and early 20th centuries. The resulting search for a

way to practice statistics without priors led to the development of frequentist statistics by such eminent figures as Sir Ronald Fisher, Karl Pearson, Jerzy Neyman, Abraham Wald, and many others.

The frequentist way of thinking came to dominate statistical theory and practice in the 20th century, to the point that most students who take only introductory statistics courses are never even aware of the existence of an alternative paradigm. However, recent decades have seen a resurgence of Bayesian statistics (partially due to advances in computing power), and an increasing number of statisticians subscribe to the Bayesian school of thought. Perhaps most encouragingly, both frequentists and Bayesians have become more willing to recognize the strengths of the opposite approach and the weaknesses of their own, and it is now common for open-minded statisticians to freely use techniques from both sides when appropriate.

## 1.1 Advantages of Bayesian Methods

The basic philosophical difference between the frequentist and Bayesian paradigms is that Bayesians treat an unknown parameter $\theta$ as *random* and use probability to quantify their uncertainty about it. In contrast, frequentists treat $\theta$ as unknown but *fixed*, and they therefore believe that probability statements about $\theta$ are useless. This fundamental disagreement leads to entirely different ways to handle statistical problems, even problems that might at first seem very basic.

To motivate the Bayesian approach, we now discuss two simple examples in which the frequentist way of thinking leads to answers that might be considered awkward, or even nonsensical.

**Example 1.1:** Let $\theta$ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Now suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, tails (5 heads, 1 tails)

To perform a frequentist hypothesis test, we must define a random variable to describe the data. The proper way to do this depends on exactly which of the following two experiments was actually performed:

- Suppose that the experiment was "Flip six times and record the results." In this case, the random variable $X$ counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$. The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned}\text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \\ &= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.109375 > 0.05.\end{aligned}$$

  So we fail to reject $H_0$ at $\alpha = 0.05$.

- Suppose instead that the experiment was "Flip until we get tails." In this case, the random variable $X$ counts the number of the flip on which the first tails occurs, and $X \sim \text{Geometric}(1 - \theta)$. The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned}\text{p-value} &= P_{\theta=1/2}(X \geq 6) \\ &= 1 - P_{\theta=1/2}(X < 6) \\ &= 1 - \sum_{x=1}^{5} P_{\theta=1/2}(X = x) \\ &= 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}\right) = \frac{1}{32} = 0.03125 < 0.05.\end{aligned}$$

  So we reject $H_0$ at $\alpha = 0.05$.

The conclusions differ, which seems absurd. Moreover the p-values aren't even close—one is 3.5 times as large as the other. Essentially, the result of our hypothesis test depends on whether we would have stopped flipping if we had gotten a tails sooner. In other words, the frequentist approach requires us to specify what we would have done had the data been something that we already know it wasn't.

Note that despite the different results, the likelihood for the actual value of $x$ that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1 - \theta).$$

A Bayesian approach would take the data into account only through this likelihood and would therefore be guaranteed to provide the same answers regardless of which experiment was being performed.

**Example 1.2:** Suppose we want to test whether the voltage $\theta$ across some electrical component differs from 9 V, based on noisy readings of this voltage from a voltmeter. Suppose the data is as follows:

9.7, 9.4, 9.8, 8.7, 8.6

A frequentist might assume that the voltage readings $X_i$ are iid from some $N(\theta, \sigma^2)$ distribution, which would lead to a basic one-sample $t$-test.

However, the frequentist is then presented with an additional piece of information: The voltmeter used for the experiment only went up to 10 V, and any readings that might have otherwise been higher are instead truncated to that value. Notice that *none of the voltages in the data are 10 V*. In other words, we already know that the 10 V limit was completely irrelevant for the data we actually observed.

Nevertheless, a frequentist must now redo the analysis and could perhaps obtain a different conclusion, because the 10 V limit changes the distribution of the observations under the null hypothesis. Like in the last example, the frequentist results change based on what would have happened had the data been something that we already know it wasn't.

The problems in Examples 1.1 and 1.2 arise from the way the frequentist paradigm forces itself to interpret probability. Another familiar aspect of this problem is the awkward definition of "confidence" in frequentist confidence intervals. The most natural interpretation of a 95% confidence interval $(L, U)$—that there is a 95% chance that the parameter is between $L$ and $U$—is dead wrong from the frequentist point of view. Instead, the notion of "confidence" must be interpreted in terms of repeating the experiment a large number of times (in principle, an infinite number), and no probabilistic statement can be made about *this particualar* confidence interval computed from the data we actually observed.

## 1.2 de Finetti's Theorem

In this section, well motivate the use of priors on parameters and indeed motivate the very use of parameters. We begin with a denition.

Definition 1.1: (Infinite exchangeability). We say that $(x_1, x_2, \dots)$ is an infinitely exchangeable sequence of random variables if, for any $n$, the joint

probability $p(x_1, x_2, ..., x_n)$ is invariant to permutation of the indices. That is, for any permutation $\pi$,

$$p(x_1, x_2, ..., x_n) = p(x_{\pi_1}, x_{\pi_2}, ..., x_{\pi_n}).$$

A key assumption of many statistical analyses is that the random variables being studied are independent and identically distributed (iid). Note that iid random variables are always infinitely exchangeable. However, infinite exchangeability is a much broader concept than being iid; an infinitely exchangeable sequence is not necessarily iid. For example, let $(x_1, x_2, \dots)$ be iid, and let $x_0$ be a non-trivial random variable independent of the rest. Then $(x_0 + x_1, x_0 + x_2, \dots)$ is infinitely exchangeable but not iid. The usefulness of infinite exchangeability lies in the following theorem.

**Theorem 1.1.** *(De Finetti). A sequence of random variables* $(x_1, x_2, \dots)$ *is infinitely exchangeable iff, for all* $n$*,*

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta),$$

*for some measure* $P$ *on* $\theta$*.*

If the distribution on $\theta$ has a density, we can replace $P(d\theta)$ with $p(\theta)\, d\theta$, but the theorem applies to a much broader class of cases than just those with a density for $\theta$.

Clearly, since $\prod_{i=1}^{n} p(x_i|\theta)$ is invariant to reordering, we have that any sequence of distributions that can be written as

$$\int \prod_{i=1}^{n} p(x_i|\theta)\; p(\theta)\; d\theta,$$

for all $n$ must be infinitely exchangeable. The other direction, though, is much deeper. It says that if we have exchangeable data, then:

- There must exist a parameter $\theta$.
- There must exist a likelihood $p(x|\theta)$.
- There must exist a distribution $P$ on $\theta$.
- The above quantities must exist so as to render the data $(x_1, \dots, x_n)$ conditionally independent.

Thus, the theorem provides an answer to the questions of why we should use parameters and why we should put priors on parameters.

**Example 1.3:** (Document processing and information retrieval). To highlight the difference between iid and infinitely exchangeable sequences, consider that search engines have historically used "bag-of-words" models to model documents. That is, for the moment, pretend that the order of words in a document does not matter. Even so, the words are definitely not iid. If we see one word and it is a French word, we then expect that the rest of the document is likely to be in French. If we see the French words *voyage* (travel), *passeport* (passport), and *douane* (customs), we expect the rest of the document to be both in French and on the subject of travel. Since we are assuming infinite exchangeability, there is some $\theta$ governing these intuitions. Thus, we see that $\theta$ can be very rich, and it seems implausible that $\theta$ might always be finite-dimensional in Theorem 2. In fact, it is the case that $\theta$ can be infinite-dimensional in Theorem 2. For example, in nonparametric Bayesian work, $\theta$ can be a stochastic process.

# Chapter 2

# Introduction to Bayesian Methods

*Every time I think I know what's going on, suddenly there's another layer of complicitions. I just want this damned thing solved.*
—John Scalzi, The Lost Colony

We introduce Bayesian methods first by motivations in decision theory and introducing the ideas of loss functions, as well as many others. Advanced topics in decision theory will be covered much later in the course. We will cover the following topics as well in this chapter:

- hierarchical and empirical Bayesian methods
- the difference between subjective and objective priors
- posterior predictive distributions.

## 2.1 Decision Theory

Another motivation for the Bayesian approach is decision theory. Its origins go back to Von Neumann and Morgenstern's game theory, but the main character was Wald. In statistical decision theory, we formalize good and bad results with a loss function.

A loss function $L(\theta, \delta(x))$ is a function of $\theta \in \Theta$ a parameter or index, and $\delta(x)$ is a decision based on the data $x \in X$. For example, $\delta(x) = n^{-1} \sum_{i=1}^{n} x_i$ might be the sample mean, and $\theta$ might be the true mean. The loss function determines the penalty for deciding $\delta(x)$ if $\theta$ is the true parameter. To give some intuition, in the discrete case, we might use a 0–1 loss, which assigns

$$L(\theta, \delta(x)) = \begin{cases} 0 & \text{if } \delta(x) = \theta, \\ 1 & \text{if } \delta(x) \neq \theta, \end{cases}$$

or in the continuous case, we might use the squared error loss $L(\theta, \delta(x)) = (\theta - \delta(x))^2$. Notice that in general, $\delta(x)$ does not necessarily have to be an estimate of $\theta$. Loss functions provide a very good foundation for statistical decision theory. They are simply a function of the state of nature ($\theta$) and a decision function ($\delta(\cdot)$). In order to compare procedures we need to calculate which procedure is best even though we cannot observe the true nature of the parameter space $\Theta$ and data $X$. This is the main challenge of decision theory and the break between frequentists and Bayesians.

## 2.2 Frequentist Risk

Definition 2.1: The frequentist risk is

$$R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))] = \int_X L(\theta, \delta) f(x|\theta)\, dx.$$

where $\theta$ is held fixed and the expectation is taken over $X$.

Thus, the risk measures the long-term average loss resulting from using $\delta$.

Figure 1 shows the risk of three different decisions as a function of $\theta \in \Theta$.

Often one decision does not dominate the other everywhere as is the case with decisions $\delta_1, \delta_2$. The challenge is in saying whether, for example, $\delta_1$ or $\delta_3$ is better. In other words, how should we aggregate over $\Theta$?

Frequentists have a few answers for deciding which is better:

1. **Admissibility**. A decision which is inadmissible is one that is dominated everywhere. For example, in Figure 1, $\delta_2$ dominates $\delta_1$ for all

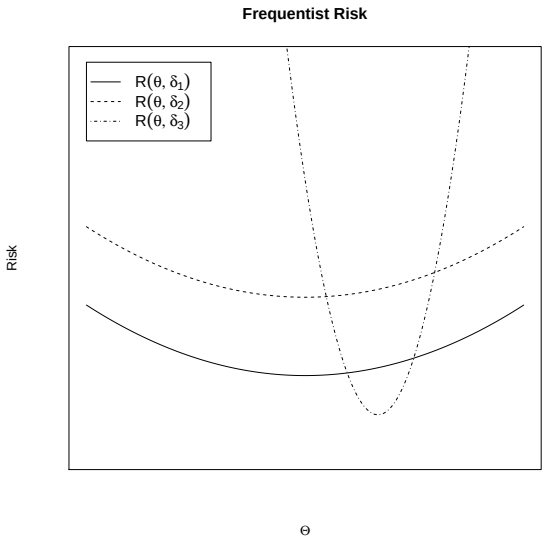



FIGURE 2.1: frequentist Risk

values of $\theta$. It would be easy to compare decisions if all but one were inadmissible. But usually the risk functions overlap, so this criterion fails.

2. **Restricted classes of procedure**. We say that an estimator $\hat{\theta}$ is an unbiased estimator of $\theta$ if $E_\theta[\hat{\theta}] = \theta$ for all $\theta$. If we restrict our attention to only unbiased estimators then we can often reduce the situation to only risk curves like $\delta_1$ and $\delta_2$ in Figure 5.21, eliminating overlapping curves like $\delta_3$. The existence of an optimal unbiased procedure is a nice frequentist theory, but many good procedures are biased—for example Bayesian procedures are typically biased. More surprisingly, some unbiased procedures are actually inadmissible. For example, James and Stein showed that the sample mean is an inadmissible estimate of the mean of a multivariate Gaussian in three or more dimensions. There are also some problems were no unbiased estimator exists—for example, when $p$ is a binomial proportion and we wish to estimate $1/p$ (see Example 2.1.2 on page 83 of Lehmann and Casella). If we restrict our class of procedures to those which are equivariant, we also get nice properties. We do not go into detail here, but these are procedures with the same group theoretic properties as the data.

3. **Minimax**. In this approach we get around the problem by just looking at $\sup_\Theta R(\theta, \delta(x))$, where $R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))]$. For example in Figure 2, $\delta_2$ would be chosen over $\delta_1$ because its maximum worst-case risk (the grey dotted line) is lower.

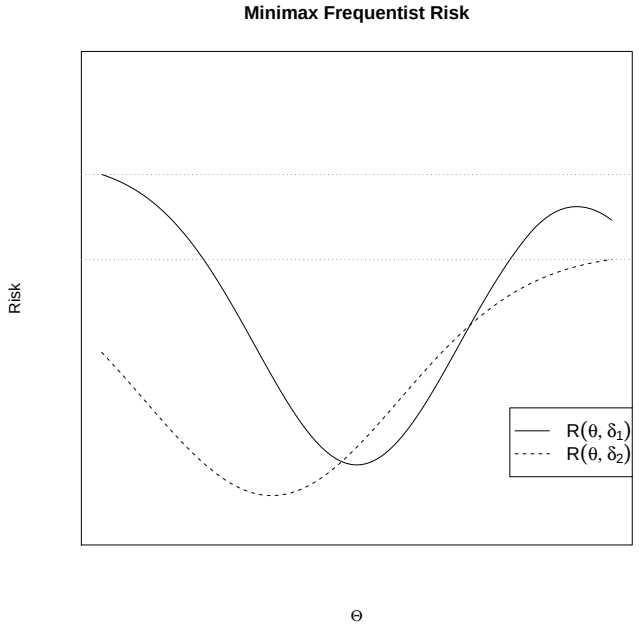



FIGURE 2.2: Minimax frequentist Risk

A Bayesian answer is to introduce a weighting function $p(\theta)$ to tell which part of $\Theta$ is important and integrate with respect to $p(\theta)$. In some sense the frequentist approach is the opposite of the Bayesian approach. However, sometimes an equivalent Bayesian procedure can be derived using a certain prior. Before moving on, again note that $R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))]$ is an expectation on $X$, assuming fixed $\theta$. A Bayesian would only look at $x$, the data you observed, not all possible $X$. An alternative definition of a frequentist is, "Someone who is happy to look at other data they could have gotten but didn't."

## 2.3 Motivation for Bayes

The Bayesian approach can also be motivated by a set of principles. Some books and classes start with a long list of axioms and principles conceived in the 1950s and 1960s. However, we will focus on three main principles.

The Bayesian approach can also be motivated by a set of principles. Some books and classes start with a long list of axioms and principles conceived in the 1950s and 1960s. However, we will focus on three main principles.

1. **Conditionality Principle**: The idea here for a Bayesian is that we *condition* on the data x.

   - Suppose we have an experiment concerning inference about $\theta$ that is chosen from a collection of possible experiments independently.
   - Then any experiment not chosen is irrelevant to the inference (this is the opposite of what we do in frequentist inference).

   **Example 2.1:** For example, two different labs estimate the potency of drugs. Both have some error or noise in their measurements which can accurately estimated from past tests. Now we introduce a new drug. Then we test its potency at a randomly chosen lab. Suppose the sample sizes matter dramatically.

   - Suppose the sample size of the first experiment (lab 1) is 1 and the sample size of the second experiment (lab 2) is 100.
   - What happens if we're doing a frequentist experiment in terms of the variance? Since this is a randomized experiment, we need to take into account all of the *data*. In essence, the variance will do some sort of averaging to take into account the sample sizes of each.
   - However, taking a Bayesian approach, we just care about the *data that we see.* Thus, the variance calculation will only come from the actual data at the randomly chosen lab.

   Thus, the question that we ask is should we use the noise level from the lab where it is tested or average over both? Intuitively, we use the noise level from the lab where it was tested, but in some frequentist approaches, it is not always so straightforward.

2. **Likelihood Principle**: The relevant information in any inference about $\theta$ after observing $x$ is contained entirely in the likelihood function. Remember the likelihood function $p(x|\theta)$ for fixed $x$ is viewed as a function of $\theta$, not $x$. For example in Bayesian approaches, $p(\theta|x) \propto p(x|\theta)p(\theta)$, so clearly inference about $\theta$ is based on the likelihood. Another approach based on the likelihood principle is Fisher's maximum likelihood estimation. This approach can also be justified by asymptotics. In case this principle seems too indisputable, here is an example using hypothesis testing in coin tossing that shows how some reasonable procedures may not follow it.

   **Example 2.2:** Let $\theta$ be the probability of a particular coin landing on heads and let

   $$H_o : \theta = 1/2,\ H_1 : \theta > 1/2.$$

   Now suppose we observe the following sequence of flips:

   $$\text{H,H,T,H,T,H,H,H,H,H,H,H,T} \qquad \text{(9 heads, 3 tails)}$$

   Then the likelihood is simply

   $$p(x|\theta) \propto \theta^9(1-\theta)^3.$$

   Many non-Bayesian analyses would pick an experimental design that is reflected in $p(x|\theta)$, for example binomial (toss a coin 12 times) or negative binomial (toss a coin until you get 3 tails). However the two lead to different probabilities over the sample space $X$. This results in different assumed tail probabilities and p-values.

   We repeat a few definitions from mathematical statistics that will be used in the course at one point or another. For example of sufficient statistics or distributions that fall in exponential families, we refer the reader to Theory of Point Estimation (TPE), Chapter 1.

   Definition 2.2: Sufficiency
   Recall that for a data set $x = (x_1, \ldots, x_n)$, a sufficient statistic $T(x)$ is a function such that the likelihood $p(x|\theta) = p(x_1, \ldots, x_n|\theta)$ depends on $x_1, \ldots, x_n$ only through $T(x)$. Then the likelihood $p(x|\theta)$ may be written as $p(x|\theta) = g(\theta, T(x))\ h(x)$ for some functions $g$ and $h$.

   Definition 2.3: Exponential Families
   A family $\{P_\theta\}$ of distributions is said to form an s-dimensional exponential family if the distributions of $P_\theta$ have densities of the form

   $$p_\theta(x) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta)T_i(x) - B(\theta)\right]h(x).$$

3. **Sufficiency Principle**: The sufficiency principle states that if two different observations $x$ and $y$ have the same sufficient statistic $T(x) = T(y)$, then inference based on $x$ and $y$ should be the same. The sufficiency principle is the least controversial principle.

   **Theorem 2.1.** *The posterior distribution, $p(\theta|y)$ only depends on the data through the sufficient statistic, T(y).*

   *Proof.* By the factorization theorem, if $T(y)$ is sufficient,

   $$f(y|\theta) = g(\theta, T(y))\ h(y).$$

   Then we know the posterior can be written

   $$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)\ d\theta} \\ &= \frac{g(\theta, T(y))\ h(y)\pi(\theta)}{\int g(\theta, T(y))\ h(y)\pi(\theta)\ d\theta} \\ &= \frac{g(\theta, T(y))\pi(\theta)}{\int g(\theta, T(y))\pi(\theta)\ d\theta} \\ &\propto g(\theta, T(y))\ p(\theta), \end{aligned}$$

   which only depends on $y$ through $T(y)$. □

   **Example 2.3:** Sufficiency
   Let $y := \sum_i y_i$. Consider

   $$y_1, \ldots, y_n|\theta \overset{iid}{\sim} \text{Bin}(1, \theta).$$

   Then $p(y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. Let p($\theta$) represent a general prior. Then

   $$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}p(\theta),$$

   which only depends on the data through the sufficient statistic $y$.

   **Theorem 2.2.** *(Birnbaum).*
   *Sufficiency Principle + Conditionality Principle = Likelihood Principle.*

   So if we assume the sufficiency principle, then the conditionality and likelihood principles are equivalent. The Bayesian approach satisfies all of these principles.

## 2.4 Bayesian Decision Theory

Earlier we discussed the frequentist approach to statistical decision theory. Now we discuss the Bayesian approach in which we condition on $x$ and integrate over $\Theta$ (remember it was the other way around in the frequentist approach). The posterior risk is defined as $\rho(\pi, \delta(x)) = \int_\Theta L(\theta, \delta(x))\pi(\theta|x)\, d\theta$.

The Bayes action $\delta^*(x)$ for any fixed $x$ is the decision $\delta(x)$ that minimizes the posterior risk. If the problem at hand is to estimate some unknown parameter $\theta$, then we typically call this the Bayes estimator instead.

**Theorem 2.3.** *Under squared error loss, the decision $\delta(x)$ that minimizes the posterior risk is the posterior mean.*

*Proof.* Suppose that $L(\theta, \delta(x)) = (\theta - \delta(x))^2$. Now note that

$$\begin{aligned}\rho(\pi, \delta(x)) &= \int (\theta - \delta(x))^2 \pi(\theta|x)\, d\theta \\ &= \int \theta^2 \pi(\theta|x)\, d\theta + \delta(x)^2 \int \pi(\theta|x)\, d\theta - 2\delta(x) \int \theta\pi(\theta|x)\, d\theta.\end{aligned}$$

Then

$$\frac{\partial[\rho(\pi, \delta(x))]}{\partial[\delta(x)]} = 2\delta(x) - 2\int \theta\pi(\theta|x)\, d\theta = 0 \iff \delta(x) = E[\theta|x],$$

and $\partial^2[\rho(\pi, \delta(x))]/\partial[\delta(x)]^2 = 2 > 0$, so $\delta(x) = E[\theta|x]$ is the minimizer. □

Recall that decision theory provides a quantification of what it means for a procedure to be 'good.' This quantification comes from the loss function $L(\theta, \delta(x))$. Frequentists and Bayesians use the loss function differently.

### ○ Frequentist Interpretation: Risk

In frequentist usage, the parameter $\theta$ is fixed, and thus it is the sample space over which averages are taken. Letting $R(\theta, \delta(x))$ denote the frequentist risk, recall that $R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))]$. This expectation is taken over the data $X$, with the parameter $\theta$ held fixed. Note that the data, $X$, is capitalized, emphasizing that it is a random variable.

**Example 2.4:** (Squared error loss). Let the loss function be squared error. In this case, the risk is

$$\begin{aligned} R(\theta, \delta(x)) &= E_\theta[(\theta - \delta(x))^2] \\ &= E_\theta[\{\theta - E_\theta[\delta(x)] + E_\theta[\delta(x)] - \delta(x)\}^2] \\ &= \{\theta - E_\theta[\delta(x)]\}^2 + E_\theta[\{\delta(x) - E_\theta[\delta(x)]\}^2 \\ &= \text{Bias}^2 + \text{Variance} \end{aligned}$$

This result allows a frequentist to analyze the variance and bias of an estimator separately, and can be used to motivate frequentist ideas, e.g. minimum variance unbiased estimators (MVUEs).

### ○ Bayesian Interpretation: Posterior Risk

Bayesians do not find the previous idea compelling because it doesn't adhere to the conditionality principle since it averages over all possible data sets. Hence, in a Bayesian framework, we define the posterior risk $\rho(x, \pi)$ based on the data $x$ and a prior $\pi$, where

$$\rho(\pi, \delta(x)) = \int_\Theta L(\theta, \delta(x))\pi(\theta|x)\, d\theta.$$

Note that the prior enters the equation when calculating the posterior density. Using the Bayes risk, we can define a bit of jargon. Recall that the Bayes action $\delta^*(x)$ is the value of $\delta(x)$ that minimizes the posterior risk. We already showed that the Bayes action under squared error loss is the posterior mean.

### ○ Hybrid Ideas

Despite the tensions between frequentists and Bayesians, they occasionally steal ideas from each other.

Definition 2.4: The Bayes risk is denoted by $r(\pi, \delta(x))$. While the Bayes risk is a frequentist concept since it averages over $X$, the expression can also

be interpreted differently. Consider

$$r(\pi, \delta(x)) = \int \int L(\theta, \delta(x))\ f(x|\theta)\ \ \pi(\theta)\ dx\ d\theta$$
$$r(\pi, \delta(x)) = \int \int L(\theta, \delta(x))\ \pi(\theta|x)\ \ \pi(x)\ dx\ d\theta$$
$$r(\pi, \delta(x)) = \int \rho(\pi, \delta(x))\ \pi(x)\ dx.$$

Note that the last equation is the posterior risk averaged over the marginal distribution of x. Another connection with frequentist theory includes that finding a Bayes rule against the "worst possible prior" gives you a minimax estimator. While a Bayesian might not find this particularly interesting, it is useful from a frequentist perspective because it provides a way to compute the minimax estimator.

We will come back to more decision theory in a more later chapter on advanced decision theory, where we will cover topics such as minimaxity, admissibility, and James-Stein estimators.

## 2.5 Bayesian Parametric Models

For now we will consider parametric models, which means that the parameter $\theta$ is a fixed-dimensional vector of numbers. Let $x \in X$ be the observed data and $\theta \in \Theta$ be the parameter. Note that $X$ may be called the sample space, while $\Theta$ may be called the parameter space. Now we define some notation that we will reuse throughout the course:

| | |
|---|---|
| $p(x\|\theta)$ | likelihood |
| $\pi(\theta)$ | prior |
| $p(x) = \int p(x\|\theta)\pi(\theta)\ d\theta$ | marginal likelihood |
| $p(\theta\|x) = \frac{p(x\|\theta)\pi(\theta)}{p(x)}$ | posterior probability |
| $p(x_{new}\|x) = \int p(x_{new}\|\theta)\pi(\theta\|x)\ d\theta$ | predictive probability |

Most of Bayesian analysis is calculating these quantities in some way or another. Note that the definition of the predictive probability assumes exchangeability, but it can easily be modied if the data are not exchangeable.

As a helpful hint, note that for the posterior distribution,

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

and oftentimes it's best to not calculate the normalizing constant $p(x)$ because you can recognize the form of $p(x|\theta)\pi(\theta)$ as a probability distribution you know. So don't normalize until the end!

**Remark**: Note that the prior distribution that we take on $\theta$ doesn't have to be a proper distribution, however, the posterior is always required to be proper for valid inference. By proper, I mean that the distribution must integrate to 1.

Two questions we still need to address are

- How do we choose priors?
- How do we compute the aforementioned quantities, such as posterior distributions?

We'll focus on choosing priors for now.

## 2.6 How to Choose Priors

We will discuss objective and subjective priors. Objective priors may be obtained from the likelihood or through some type of invariance argument. Subjective priors are typically arrived at by a process involving interviews with domain experts and thinking really hard; in fact, there is arguably more philosophy and psychology in the study of subjective priors than mathematics. We start with conjugate priors. The main justification for the use of conjugate priors is that they are computationally convenient and they have asymptotically desirably properties.

Choosing prior probabilities: Subjective or Objective.

**Subjective**
A prior probability could be subjective based on the information a person

might have due to past experience, scientific considerations, or simple common sense. For example, suppose we wish to estimate the probability that a randomly selected woman has breast cancer. A simple prior could be formulated based on the national or worldwide incidence of breast cancer. A more sophisticated approach might take into account the woman's age, ethnicity, and family history. Neither approach could necessarily be classified as right or wrong—again, it's subjective.

As another example, say a doctor administers a treatment to patients and finds 48 out of 50 are cured. If the same doctor later wishes to investigate the cure probability of a similar but slightly different treatment, he might expect that its cure probability will also be around $48/50 = 0.96$. However a different doctor may have only had 8/10 patients cured by the first treatment and might therefore specify a prior suggesting a cure rate of around 0.8 for the for the new treatment. For convenience, subjective priors are often chosen to take the form of common distributions, such as the normal, gamma, or beta distribution.

**Objective**
An objective prior (also called default, vague, noninformative) can also be used in a given situation even in the absence of enough information. Examples of objective priors are flat priors such as Laplace's, Haldane's, Jeffreys', and Bernardo's references priors. These priors will be discussed later.

## 2.7 Hierarchical Bayesian Models

In a hierarchical Bayesian model, rather than specifying the prior distribution as a single function, we specify it as a hierarchy. Thus, on the unknown parameter of interest, say $\theta$, we put a prior. On any other unknown *hyperparameters* of the model that are given, we also specify priors for these. We write

$$\begin{aligned} X|\theta &\sim f(x|\theta) \\ \Theta|\gamma &\sim \pi(\theta|\gamma) \\ \Gamma &\sim \phi(\gamma), \end{aligned}$$

where we assume that $\phi(\gamma)$ is known and not dependent on any other unknown hyperparameters (what the parameters of the prior are often called

as we have already said). Note that we can continue this hierarchical modeling and add more stages to the model, however note that doing so adds more complexity to the model (and possibly as we will see may result in a posterior that we cannot compute without the aid of numerical integration or MCMC, which we will cover in detail in a later chapter).

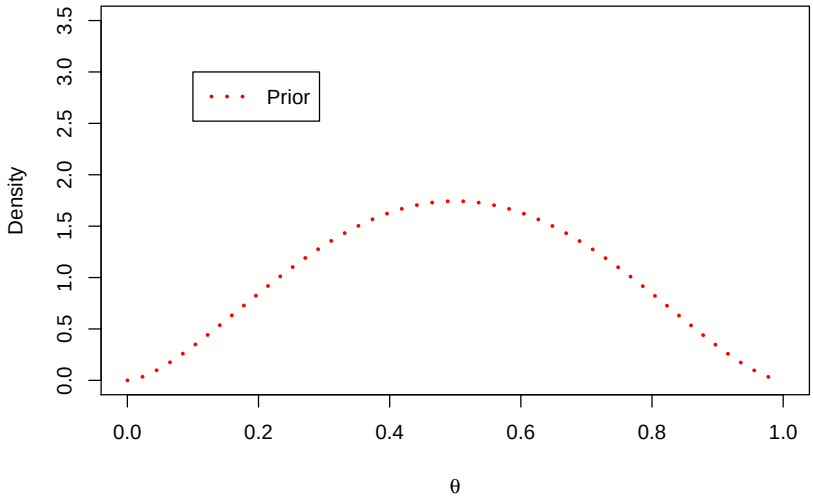
Definition 2.5: (Conjugate Distributions). Let $F$ be the class of sampling distributions $p(y|\theta)$. Then let $P$ denote the class of prior distributions on $\theta$. Then $P$ is said to be conjugate to $F$ if for every $p(\theta) \in P$ and $p(y|\theta) \in F$, $p(y|\theta) \in P$. **Simple definition**: A family of priors such that, upon being multiplied by the likelihood, yields a posterior in the same family.

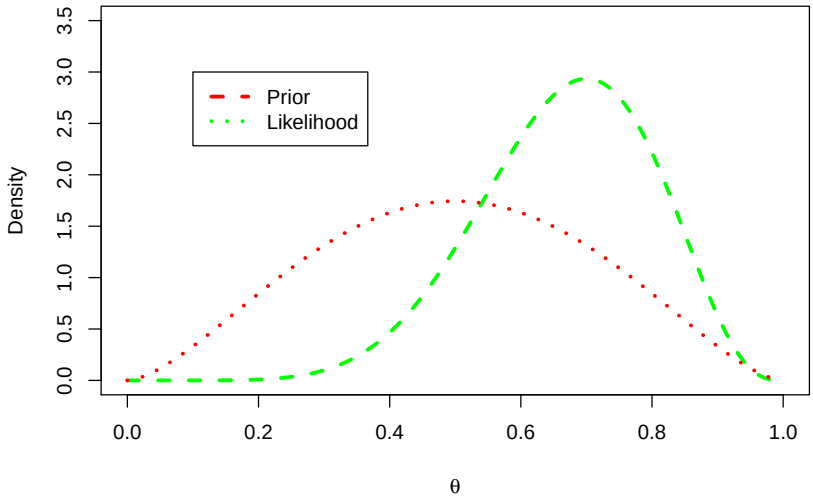
**Example 2.5:** (Beta-Binomial) If $X|\theta$ is distributed as binomial$(n, \theta)$, then a conjugate prior is the beta family of distributions, where we can show that the posterior is

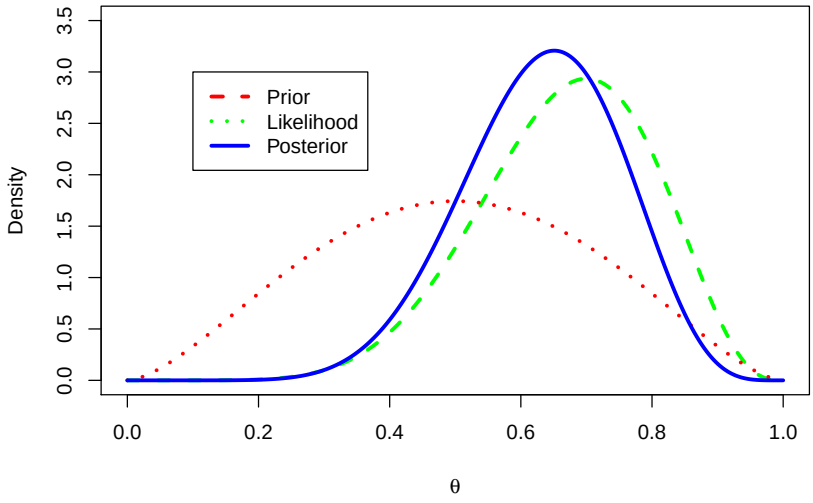
$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \implies \\
&\qquad \theta|x \sim \text{Beta}(x+a, n-x+b).
\end{aligned}$$

Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-December was 45%. We'll assume that in MA his approval rating is approximately 50%.
- Based on this prior information, we'll use a Beta prior for $\theta$ and we'll choose $a$ and $b$. (Won't get into this here).
- We can plot the prior and likelihood distributions in `R` and then see how the two mix to form the posterior distribution.


Prior
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ


Prior
Likelihood
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ




**Example 2.6:** (Normal-Uniform Prior)

$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} \text{Normal}(\theta, \sigma^2), \ \sigma^2 \text{ known}$$
$$\theta \sim \text{Uniform}(-\infty, \infty),$$

where $\theta \sim \text{Uniform}(-\infty, \infty)$ means that $p(\theta) \propto 1$.

Calculate the posterior distribution of $\theta$ given the data.

$$\begin{aligned} p(\theta|x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(x_i - \theta)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\} \\ &\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_{i}(x_i - \theta)^2\right\}. \end{aligned}$$

Note that $\sum_i (x_i - \theta)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2$. Then

$$\begin{aligned} p(\theta|x) &\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i - \bar{x})^2\right\} \exp\left\{\frac{-n}{2\sigma^2}(\bar{x} - \theta)^2\right\} \\ &\propto \exp\left\{\frac{-n}{2\sigma^2}(\bar{x} - \theta)^2\right\} \\ &= \exp\left\{\frac{-n}{2\sigma^2}(\theta - \bar{x})^2\right\}. \end{aligned}$$

Thus,

$$\theta|x_1, \ldots, x_n \sim \text{Normal}(\bar{x}, \sigma^2/n).$$

**Example 2.7:** Normal-Normal

$$\begin{aligned} X_1, \ldots, X_n|\theta &\overset{iid}{\sim} \text{N}(\theta, \sigma^2) \\ \theta &\sim \text{N}(\mu, \tau^2), \end{aligned}$$

where $\sigma^2$ is known. Calculate the distribution of $\theta|x_1, \ldots, x_n$.

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(x_i - \theta)^2\right\} \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{\frac{-1}{2\tau^2}(\theta - \mu)^2\right\} \\ &\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i - \theta)^2\right\} \exp\left\{\frac{-1}{2\tau^2}(\theta - \mu)^2\right\}. \end{aligned}$$

Consider

$$\sum_i (x_i - \theta)^2 = \sum_i (x_i - \bar{x} + \bar{x} - \theta)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2.$$

Then

$$\begin{aligned}
p(\theta|x_1,\ldots,x_n) &= \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i-\bar{x})^2\right\} \times \exp\left\{\frac{-1}{2\sigma^2}n(\bar{x}-\theta)^2\right\} \times \exp\left\{\frac{-1}{2\tau^2}(\theta-\mu)^2\right\} \\
&\propto \exp\left\{\frac{-1}{2\sigma^2}n(\bar{x}-\theta)^2\right\}\exp\left\{\frac{-1}{2\tau^2}(\theta-\mu)^2\right\} \\
&= \exp\left\{\frac{-1}{2}\left[\frac{n}{\sigma^2}\left(\bar{x}^2-2\bar{x}\theta+\theta^2\right)+\frac{1}{\tau^2}\left(\theta^2-2\theta\mu+\mu^2\right)\right]\right\} \\
&= \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\theta^2-2\theta\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}\right)+\frac{n\bar{x}^2}{\sigma^2}+\frac{\mu^2}{\tau^2}\right]\right\} \\
&\propto \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta^2-2\theta\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)\right]\right\} \\
&\propto \exp\left\{\frac{-1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta-\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2\right]\right\}.
\end{aligned}$$

Recall what it means to complete the square as we did above.[1] Thus,

$$\begin{aligned}
\theta|x_1,\ldots,x_n &\sim N\left(\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right) \\
&= N\left(\frac{n\bar{x}\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right).
\end{aligned}$$

DEFINITION 2.6: The reciprocal of the variance is referred to as the *precision*. That is,

$$\text{Precision} = \frac{1}{\text{Variance}}.$$

**Theorem 2.4.** *Let $\delta_n$ be a sequence of estimators of $g(\theta)$ with mean squared error $E(\delta_n - g(\theta))^2$.*

*(i) If $E[\delta_n - g(\theta)]^2 \to 0$ then $\delta_n$ is consistent for $g(\theta)$.*

*(ii) Equivalent to the above, $\delta_n$ is consistent if $b_n(\theta) \to 0$ and $Var(\delta_n) \to 0$ for all $\theta$.*

[1] Recall from algebra that $(x-b)^2 = x^2 - 2bx + b^2$. We want to complete something that resembles $x^2 - 2bx = x^2 + 2bx + (2b/2)^2 - (2b/2)^2 = (x-b)^2 - b^2$.

*(iii) In particular (and most useful), $\delta_n$ is consistent if it is unbiased for each $n$ and if $Var(\delta_n) \to 0$ for all $\theta$.*

*We omit the proof since it requires Chebychev's Inequality along with a bit of probability theory. See Problem 1.8.1 in TPE for the exercise of proving this.*

**Example 2.8:** (Normal-Normal Revisited) Recall Example 2.7. We write the posterior mean as $E(\theta|x)$. Let's write the posterior mean in this example as

$$\begin{aligned} E(\theta|x) &= \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}. \\ &= \frac{\frac{n\bar{x}}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \frac{\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}. \end{aligned}$$

We also write the posterior variance as

$$V(\theta|x) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

We can see that the posterior mean is a weighted average of the sample mean and the prior mean. The weights are proportional to the reciprocal of the respective variances (precision). In this case,

$$\begin{aligned} \text{Posterior Precision} &= \frac{1}{\text{Posterior Variance}} \\ &= (n/\sigma^2) + (1/\tau^2) \\ &= \text{Sample Precision} + \text{Prior Precision}. \end{aligned}$$

The posterior precision is larger than either the sample precision or the prior precision. Equivalently, the posterior variance, denoted by $V(\theta|x)$, is smaller than either the sample variance or the prior variance.

What happens as $n \to \infty$?

Divide the posterior mean (numerator and denominator) by $n$. Now take $n \to \infty$. Then

$$E(\theta|x) = \frac{\frac{1}{n}\frac{n\bar{x}}{\sigma^2} + \frac{1}{n}\frac{\mu}{\tau^2}}{\frac{1}{n}\frac{n}{\sigma^2} + \frac{1}{n}\frac{1}{\tau^2}} \to \frac{\frac{\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2}} = \bar{x} \quad \text{as } n \to \infty.$$

In the case of the posterior variance, divide the denominator and numerator by $n$. Then

$$V(\theta|x) = \frac{\frac{1}{n}}{\frac{1}{n}\frac{n}{\sigma^2} + \frac{1}{n}\frac{1}{\tau^2}} \approx \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty.$$

Since the posterior mean is unbiased and the posterior variance goes to 0, the posterior mean is consistent by Theorem 2.4.

**Example 2.9:**

$$\begin{aligned} X|\alpha,\beta &\sim \text{Gamma}(\alpha,\beta),\ \alpha \text{ known},\ \beta \text{ unknown} \\ \beta &\sim \text{IG}(a,b). \end{aligned}$$

Calculate the posterior distribution of $\beta|x$.

$$\begin{aligned} p(\beta|x) &\propto \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \times \frac{b^a}{\Gamma(a)} \beta^{-a-1} e^{-b/\beta} \\ &\propto \frac{1}{\beta^{\alpha}} e^{-x/\beta} \beta^{-a-1} e^{-b/\beta} \\ &= \beta^{-\alpha-a-1} e^{-(x+b)/\beta}. \end{aligned}$$

Notice that this looks like an Inverse Gamma distribution with parameters $\alpha + a$ and $x + b$. Thus,

$$\beta|x \sim IG(\alpha + a, x + b).$$

**Example 2.10:** (Bayesian versus frequentist)
Suppose a child is given an IQ test and his score is $X$. We assume that

$$\begin{aligned} X|\theta &\sim \text{Normal}(\theta, 100) \\ \theta &\sim \text{Normal}(100, 225) \end{aligned}$$

From previous calculations, we know that the posterior is

$$\theta|x \sim \text{Normal}\left(\frac{400 + 9x}{13}, \frac{900}{13}\right).$$

Here the posterior mean is $(400 + 9x)/13$. Suppose $x = 115$. Then the posterior mean becomes 110.4. Contrasting this, we know that the frequentist estimate is the mle, which is $x = 115$ in this example.

The posterior variance is $900/13 = 69.23$, whereas the variance of the data is $\sigma^2 = 100$.

Now suppose we take the Uniform$(-\infty, \infty)$ prior on $\theta$. From an earlier example, we found that the posterior is

$$\theta|x \sim \text{Normal}\left(115, 100\right).$$

Notice that the posterior mean and mle are both 115 and the posterior variance and variance of the data are both 100.

When we put little/no prior information on $\theta$, the data washes away most/all of the prior information (and the results of frequentist and Bayesian estimation are similar or equivalent in this case).

**Example 2.11:** (Normal Example with Unknown Variance)
Consider

$$X_1, \ldots, X_n|\theta \overset{iid}{\sim} \text{Normal}(\theta, \sigma^2),\ \theta \text{ known},\ \sigma^2 \text{ unknown}$$
$$p(\sigma^2) \propto (\sigma^2)^{-1}.$$

Calculate $p(\sigma^2|x_1, \ldots, x_n)$.

$$\begin{aligned} p(\sigma^2|x_1, \ldots, x_n) &\propto (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i - \theta)^2\right\} (\sigma^2)^{-1} \\ &\propto (\sigma^2)^{-n/2-1} \exp\left\{\frac{-1}{2\sigma^2}\sum_i (x_i - \theta)^2\right\}. \end{aligned}$$

Recall, if $Y \sim \text{IG}(a, b)$, then $f(y) = \dfrac{b^a}{\Gamma(a)} y^{-a-1} e^{-b/y}$. Thus,

$$\sigma^2|x_1, \ldots, x_n \sim \text{IG}\left(n/2, \frac{\sum_{i=1}^n (x_i - \theta)^2}{2}\right).$$

**Example 2.12:** (Football Data) Gelman et. al (2003) consider the problem of estimating an unknown variance using American football scores. The focus is on the difference $d$ between a game outcome (winning score minus losing score) and a published point spread.

- We observe $d_1, \ldots, d_n$, the observed differences between game outcomes and point spreads for $n = 2240$ football games.
- We assume these differences are a random sample from a Normal distribution with mean 0 and unknown variance $\sigma^2$.
- Our goal is to make inference on the unknown parameter $\sigma^2$, which represents the variability in the game outcomes and point spreads.

We can refer to Example 2.11, since the setup here is the same. Hence the posterior becomes

$$\sigma^2|d_1, \ldots, d_n \sim \text{IG}(n/2, \sum_i d_i^2/2).$$

The next logical step would be plotting the posterior distribution in R. As far as I can tell, there is not a built-in function predefined in R for the Inverse Gamma density. However, someone saw the need for it and built one in using the pscl package.

Proceeding below, we try and calculate the posterior using the function densigamma, which corresponds to the Inverse Gamma density. However, running this line in the code gives the following error:

```
Warning message:
In densigamma(sigmas, n/2, sum(d^2)/2) : value out of range in 'gammafn'
```

What's the problem? Think about the what the posterior looks like. Recall that

$$p(\sigma^2|\boldsymbol{d}) = \frac{(\sum_i d_i^2/2)^{n/2}}{\Gamma(n/2)}(\sigma^2)^{-n/2-1}e^{-(\sum_i d_i^2)/2\sigma^2}.$$

In the calculation R is doing, it's dividing by $\Gamma(1120)$, which is a very large factorial. This is too large for even R to compute, so we're out of luck here. So, what can we do to analyze the data?

```
setwd("~/Desktop/sta4930/football")
data = read.table("football.txt",header=T)
names(data)
attach(data)
score = favorite-underdog
d = score-spread
n = length(d)
hist(d)
install.packages("pscl",repos="http://cran.opensourceresources.org")
library(pscl)
?densigamma
sigmas = seq(10,20,by=0.1)
post = densigamma(sigmas,n/2,sum(d^2)/2)
v = sum(d^2)
```

We know we can't use the Inverse Gamma density (because of the function in R), but we do know a relationship regarding the Inverse Gamma and Gamma distributions. So, let's apply this fact.

You may be thinking, we're going to run into the same problem because we'll still be dividing by $\Gamma(1120)$. This is true, except the Gamma density function `dgamma` was built into `R` by the original writers. The `dgamma` function is able to do some internal tricks that let it calculate the gamma density even though the individual piece $\Gamma(n/2)$ by itself is too large for `R` to handle. So, moving forward, we will apply the following fact that we already learned:

$$\text{If } X \sim \text{IG}(a, b), \text{ then } 1/X \sim \text{Gamma}(a, 1/b).$$

Since

$$\sigma^2 | d_1, \ldots, d_n \sim \text{IG}(n/2, \sum_i d_i^2/2),$$

we know that

$$\frac{1}{\sigma^2} | d_1, \ldots, d_n \sim \text{Gamma}(n/2, 2/v), \quad \text{where } v = \sum_i d_i^2.$$

Now we can plot this posterior distribution in `R`, however in terms of making inference about $\sigma^2$, this isn't going to be very useful.

In the code below, we plot the posterior of $\frac{1}{\sigma^2}|\boldsymbol{d}$. In order to do so, we must create a new sequence of $x$-values since the mean of our gamma will be at $n/v \approx 0.0053$.

```
xnew = seq(0.004,0.007,.000001)
pdf("football_sigmainv.pdf", width = 5, height = 4.5)
post.d = dgamma(xnew,n/2,scale = 2/v)
plot(xnew,post.d, type= "l", xlab = expression(1/sigma^2), ylab= "density")
dev.off()
```

As we can see from the plot below, viewing the posterior of $\frac{1}{\sigma^2}|\boldsymbol{d}$ isn't very useful. We would like to get the parameter in terms of $\sigma^2$, so that we could plot the posterior distribution of interest as well as calculate the posterior mean and variance.

To recap, we know

$$\frac{1}{\sigma^2} | d_1, \ldots, d_n \sim \text{Gamma}(n/2, 2/v), \quad \text{where } v = \sum_i d_i^2.$$

Let $u = \frac{1}{\sigma^2}$. We are going to make a transformation of variables now to write the density in terms of $\sigma^2$.

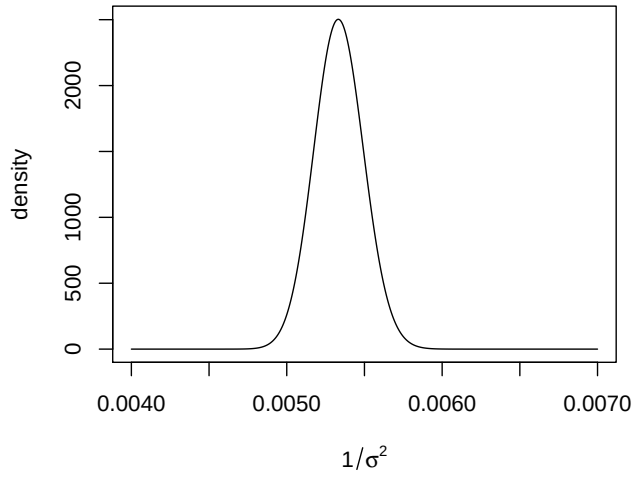



FIGURE 2.3: Posterior Distribution $p(\frac{1}{\sigma^2}|d_1, \ldots, d_n)$

Since $u = \frac{1}{\sigma^2}$, this implies $\sigma^2 = \frac{1}{u}$. Then $\left|\frac{\partial u}{\partial \sigma^2}\right| = \frac{1}{\sigma^4}$.

Now applying the transformation of variables we find that

$$f(\sigma^2|d_1, \ldots, d_n) = \frac{1}{\Gamma(n/2)(2/v)^{n/2}} \left(\frac{1}{\sigma^2}\right)^{n/2-1} e^{-\frac{v}{2\sigma^2}} \left(\frac{1}{\sigma^4}\right).$$

Thus,

$$\sigma^2|\boldsymbol{d} \sim \text{Gamma}(n/2, 2/v) \left(\frac{1}{\sigma^4}\right).$$

Now, we know the density of $\sigma^2|\boldsymbol{d}$ in a form we can calculate in R.

```
x.s = seq(150,250,1)
pdf("football_sigma.pdf", height = 5, width = 4.5)
post.s = dgamma(1/x.s,n/2, scale = 2/v)*(1/x.s^2)
plot(x.s,post.s, type="l", xlab = expression(sigma^2), ylab="density")
dev.off()
detach(data)
```

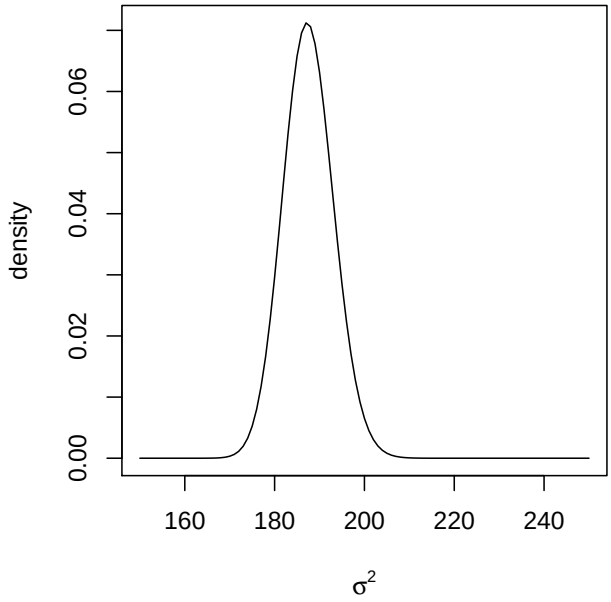



FIGURE 2.4: Posterior Distribution $p(\sigma^2|d_1, \ldots, d_n)$

From the posterior plot in Figure 2.4 we can see that the posterior mean is around 185. This means that the variability of the actual game result around the point spread has a standard deviation around 14 points. If you wanted to actually calculate the posterior mean and variance, you could do this using a numerical method in `R`.

What's interesting about this example is that there is a lot more variability in football games than the average person would most likely think.

- Assume that (1) the standard deviation actually is 14 points, and (2) game result is normally distributed (which it's not, exactly, but this is a reasonable approximation).
- Things with a normal distribution fall two or more standard deviations from their mean about 5% of the time, so this means that, roughly speaking, about 5% of football games end up 28 or more points away from their spread.

**Example 2.13:**

$$\begin{aligned} Y_1, \ldots, Y_n | \mu, \sigma^2 &\overset{iid}{\sim} \text{Normal}(\mu, \sigma^2), \\ \mu | \sigma^2 &\sim \text{Normal}(\mu_0, \frac{\sigma^2}{\kappa_0}), \\ \sigma^2 &\sim \text{IG}(\frac{\nu_0}{2}, \frac{\sigma_0^2}{2}), \end{aligned}$$

where $\mu_0, \kappa_0, \nu_0, \sigma_0^2$ are constant.

Find $p(\mu, \sigma^2 | y_1 \ldots, y_n)$. Notice that

$$\begin{aligned} p(\mu, \sigma^2 | y_1 \ldots, y_n) &= \frac{p(\mu, \sigma^2, y_1 \ldots, y_n)}{p(y_1 \ldots, y_n)} \\ &\propto p(y_1 \ldots, y_n | \mu, \sigma^2) p(\mu, \sigma^2) \\ &= p(y_1 \ldots, y_n | \mu, \sigma^2) p(\mu | \sigma^2) p(\sigma^2). \end{aligned}$$

Then

$$\begin{aligned} p(\mu, \sigma^2 | y_1 \ldots, y_n) &\propto p(y_1 \ldots, y_n | \mu, \sigma^2) p(\mu | \sigma^2) p(\sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} (\sigma^2)^{-1/2} \exp\left\{ \frac{-\kappa_0}{2\sigma^2} (\mu - \mu_0)^2 \right\} \\ &\times (\sigma^2)^{-\nu_0/2 - 1} \exp\left\{ \frac{-\sigma_0^2}{2\sigma^2} \right\}. \end{aligned}$$

Consider $\sum_i (y_i - \mu)^2 = \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$.

Then

$$\begin{aligned} n(\bar{y} - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 &= n\bar{y}^2 - 2n\bar{y}\mu + n\mu^2 + \kappa_0 \mu^2 - 2\kappa_0 \mu \mu_0 + \kappa_0 {\mu_0}^2 \\ &= (n + \kappa_0)\mu^2 - 2(n\bar{y} + \kappa_0 \mu_0)\mu + n\bar{y}^2 + \kappa_0 {\mu_0}^2 \\ &= (n + \kappa_0) \left( \mu - \frac{n\bar{y} + \kappa_0 \mu_0}{n + \kappa_0} \right)^2 - \frac{(n\bar{y} + \kappa_0 \mu_0)^2}{n + \kappa_0} + n\bar{y}^2 + \kappa_0 {\mu_0}^2. \end{aligned}$$

Now consider

$$\begin{aligned}
n\bar{y}^2 + \kappa_0{\mu_0}^2 - \frac{(n\bar{y}+\kappa_0\mu_0)^2}{n+\kappa_0} &= n\bar{y}^2 + \kappa_0{\mu_0}^2 + \frac{-n^2\bar{y}^2 - 2n\kappa_0\mu_0\bar{y} - {\kappa_0}^2{\mu_0}^2}{n+\kappa_0} \\
&= \frac{n^2\bar{y}^2 + n\kappa_0{\mu_0}^2 + n\kappa_0\bar{y}^2 + {\kappa_0}^2{\mu_0}^2 - n^2\bar{y}^2 - 2n\kappa_0\mu_0\bar{y} - {\kappa_0}^2{\mu_0}^2}{n+\kappa_0} \\
&= \frac{n\kappa_0{\mu_0}^2 + n\kappa_0\bar{y}^2 - 2n\kappa_0\mu_0\bar{y}}{n+\kappa_0} \\
&= \frac{n\kappa_0({\mu_0}^2 - 2\mu_0\bar{y} + \bar{y}^2)}{n+\kappa_0} \\
&= \frac{n\kappa_0(\mu_0-\bar{y})^2}{n+\kappa_0}.
\end{aligned}$$

Putting this all together, we find

$$\begin{aligned}
p(\mu,\sigma^2|y_1\ldots,y_n) &\propto \exp\left\{\frac{-n}{2\sigma^2}(\bar{y}-\mu)^2\right\}\exp\left\{\frac{-1}{2\sigma^2}\sum_i(y_i-\bar{y})^2\right\} \\
&\quad\times\exp\left\{\frac{-\kappa_0}{2\sigma^2}\sum_i(\mu-\mu_0)^2\right\}(\sigma^2)^{-n/2-1/2}(\sigma^2)^{-\nu_0/2-1}\exp\left\{\frac{-\sigma_0^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{-n\kappa_0}{2\sigma^2(n+\kappa_0)}(\mu_0-\bar{y})^2\right\}\exp\left\{\frac{-1}{2\sigma^2}\sum_i(y_i-\bar{y})^2\right\} \\
&\quad\times\exp\left\{-\frac{(n+\kappa_0)}{2\sigma^2}\left(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right)^2\right\}(\sigma^2)^{-\nu_0/2-1}(\sigma^2)^{-n/2-1}\exp\left\{\frac{-\sigma_0^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{-1}{2\sigma^2}\sum_i(y_i-\bar{y})^2 - \frac{n\kappa_0}{2\sigma^2(n+\kappa_0)}(\mu_0-\bar{y})^2 - \frac{\sigma_0^2}{2\sigma^2}\right\}(\sigma^2)^{-(n+\nu_0)/2-1} \\
&\quad\times\exp\left\{-\frac{(n+\kappa_0)}{2\sigma^2}\left(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right)^2\right\}(\sigma^2)^{-1/2}.
\end{aligned}$$

Since the posterior above factors, we find

$$\mu|\sigma^2,\boldsymbol{y} \sim \text{Normal}\left(\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}, \frac{\sigma^2}{n+\kappa_0}\right),$$

$$\sigma^2|\boldsymbol{y} \sim IG\left(\frac{n+\nu_0}{2}, \frac{1}{2}\left(\sum_i(y_i-\bar{y})^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\mu_0-\bar{y})^2 + \sigma_0^2\right)\right).$$

**Example 2.14:** Suppose we calculate $E[\theta|y]$ where $y = x_{(n)}$. Let

$$\begin{aligned}
X_i \mid \theta &\sim \text{Uniform}(0,\theta) \\
\theta &\sim \text{Gamma}(a, 1/b).
\end{aligned}$$

Show

$$E[\theta|x] = \frac{1}{b(n+a-1)} \frac{P(\chi^2_{2(n+a-1)} < 2/(by))}{P(\chi^2_{2(n+a-1)} < 2/(by))}.$$

*Proof.* Recall that the posterior depends on the data only through the sufficient statistic $y$. Consider that $P(Y \le y) = P(X_1 \le y)^n = (y/\theta)^n \implies f_y(y) = n/\theta(y/\theta)^{n-1} = \frac{n}{\theta^n} y^{n-1}$.

$$\begin{aligned} E[\theta|x] &= \frac{\int \theta f(y|\theta)\pi(\theta)\, d\theta}{\int f(y|\theta)\pi(\theta)\, d\theta} \\ &= \frac{\int_y^\infty \frac{\theta\, n\, y^{n-1}\theta^{-a-1}e^{-1/(\theta b)}}{\theta^n \Gamma(a) b^a\, d\theta}}{\int_y^\infty \frac{n\theta^{a-1}e^{-1/(\theta b)}}{\theta^n\Gamma(a)b^a\, d\theta}} \\ &= \frac{\int_y^\infty \theta^{-n-a}e^{-1/(\theta b)}\, d\theta}{\int_y^\infty \theta^{-n-a-1}e^{-1/(\theta b)}\, d\theta} \end{aligned}$$

Let $\theta = 2/(xb) \implies d\theta = -2/(bx^2)\, dx$. Recall that Gamma$(v/2, 2)$ is a $\chi^2_v$. Then

$$\begin{aligned} E[\theta|x] &= \frac{\int_0^{\frac{2}{by}} (\frac{2}{xb})^{-n-a} e^{-x/2} \frac{2}{bx^2}\, dx}{\int_0^{\frac{2}{by}} (\frac{2}{xb})^{-n-a-1} e^{-x/2} \frac{2}{bx^2}\, dx} \\ &= \frac{\frac{\int_0^{\frac{2}{by}} b^{n+a-1}x^{n+a-2}e^{-x/2}\, dx \times \Gamma(n+a-1)\, dx}{2^{n+a-1}\Gamma(n+a-1)}}{\frac{\int_0^{\frac{2}{by}} b^{n+a+1-1}x^{n+a+1-2}e^{-x/2}\, dx \times \Gamma(n+a)\, dx}{2^{n+a+1-1}\Gamma(n+a)}} \\ &= \frac{P(\chi^2_{2(n+a-1)} < 2/(by))}{P(\chi^2_{2(n+a-1)} < 2/(by))} \frac{b^{n+a-1}\Gamma(n+a-1)}{b^{n+a}\Gamma(n+a)} \\ &= \frac{1}{b(n+a-1)} \frac{P(\chi^2_{2(n+a-1)} < 2/(by))}{P(\chi^2_{2(n+a-1)} < 2/(by))}. \end{aligned}$$

□

## 2.8 Empirical Bayesian Models

Another generalization of Bayes estimation is called empirical Bayes (EB) estimation, which most consider to fall outside of the Bayesian paradigm (in the sense that it's not fully Bayesian). However, it's been proved to be a technique of constructing estimators that perform well under both Bayesian and frequentist criteria. One reason for this is that EB estimators tend to be more robust against model misspecification of the prior distribution.

We start again with an HB model, however this time we assume that $\gamma$ is unknown and must be estimated. We begin with the Bayes model

$$\begin{aligned} X_i|\theta &\sim f(x|\theta),\ i = 1\ldots,p \\ \Theta|\gamma &\sim \pi(\theta|\gamma). \end{aligned}$$

We then calculate the marginal distribution of $\boldsymbol{X}$ with density

$$m(\boldsymbol{x}|\gamma) = \int \prod f(x_i|\theta)\pi(\theta|\gamma)\ d\theta.$$

Based on $m(\boldsymbol{x}|\gamma)$, we obtain an estimate of $\hat{\gamma}(x)$ of $\gamma$. It's most common to find the estimate using maximum likelihood estimation (MLE), but method of moments could be used as well (or other methods). We now substitute $\hat{\gamma}(x)$ for $\gamma$ in $\pi(\theta|\gamma)$ and determine the estimator that minimizes the empirical posterior loss

$$\int L(\theta,\delta)\pi(\theta|\hat{\gamma}(x))\ d\theta.$$

**Remark**: An alternative definition is obtained by substituting $\hat{\gamma}(x)$ for $\gamma$ in the Bayes estimator. (This proof is left as a homework exercise, 4.6.1 in TPE).

**Example 2.15:** Empirical Bayes Binomial
Suppose there are $K$ different groups of patients where each group has $n$ patients. Each group is given a different treatment for the same illness and in the $k$th group, we count $X_k$, $k = 1,\ldots,K$, which is the number of successful treatments our of $n$.

Since the groups receive different treatments, we expect different success rates, however, since we are treating the same illness, these rates should be related to each other. These considerations suggest the following model:

$$X_k \sim \text{Bin}(n, p_k),$$
$$p_k \sim \text{Beta}(a, b),$$

where the $K$ groups are tied together by the common prior distribution.

It is easy to show that the Bayes estimator of $p_k$ under squared error loss is

$$E(p_k|a_k, a, b) = \frac{a + x_k}{a + b + n}.$$

Suppose now that we are told that $a, b$ are unknown and we wish to estimate them using EB. We first calculate

$$\begin{aligned} m(\boldsymbol{x}|a, b) &= \int_{0,1} \cdots \int_{0,1} \prod_{k=1}^{K} \binom{n}{x_k} p_k^{x_k} (1 - p_k)^{n - x_k} \times \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p_k^{a-1} (1 - p_k)^{b-1} \, dp_k \\ &= \int_{0,1} \cdots \int_{0,1} \prod_{k=1}^{K} \binom{n}{x_k} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p_k^{x_k + a - 1} (1 - p_k)^{n - x_k + b - 1} \, dp_k \\ &= \prod_{k=1}^{K} \binom{n}{x_k} \frac{\Gamma(a + b)\Gamma(a + x_k)\Gamma(n - x_k + b)}{\Gamma(a)\Gamma(b)\Gamma(a + b + n)} \end{aligned}$$

which is a product of beta-binomials. Although the MLEs of $a$ and $b$ aren't expressible in closed form, they can be calculated numerically to construct the EB estimator

$$\hat{\delta}^{EB}(x) = \frac{\hat{a} + x_k}{\hat{a} + \hat{b} + n}.$$

## 2.9 Posterior Predictive Distributions

We have just gone through many examples illustrating how to calculate many simple posterior distributions. This is the main goal of a Bayesian analysis. Another goal might be prediction. That is given some data $y$ and a new observation $\tilde{y}$, we may wish to find the conditional distribution of $\tilde{y}$ given $y$. This distribution is referred to as the *posterior predictive distribution*. That is, our goal is to find $p(\tilde{y}|y)$. This minimizing estimator is called the empirical Bayes estimator.

We'll derive the posterior predictive distribution for the discrete case ($\theta$ is discrete). It's the same for the continuous case, with the sums replaced with integrals.

Consider

$$\begin{aligned} p(\tilde{y}|y) &= \frac{p(\tilde{y}, y)}{p(y)} \\ &= \frac{\int_\theta p(\tilde{y}, y, \theta)\, d\theta}{p(y)} \\ &= \frac{\int_\theta p(\tilde{y}|y, \theta) p(y, \theta)\, d\theta}{p(y)} \\ &= \int_\theta p(\tilde{y}|y, \theta) p(\theta|y)\, d\theta. \end{aligned}$$

In most contexts, if $\theta$ is given, then $\tilde{y}|\theta$ is independent of $y$, i.e., the value of $\theta$ determines the distribution of $\tilde{y}$, without needing to also know $y$. When this is the case, we say that $\tilde{y}$ and $y$ are *conditionally independent* given $\theta$. Then the above becomes

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

**Theorem 2.5.** *If $\theta$ is discrete and $\tilde{y}$ and $y$ are conditionally independent given $\theta$, then the posterior predictive distribution is*

$$p(\tilde{y}|y) = \sum_\theta p(\tilde{y}|\theta) p(\theta|y).$$

If $\theta$ is continuous and $\tilde{y}$ and $y$ are conditionally independent given $\theta$, then the posterior predictive distribution is

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

**Theorem 2.6.** *Suppose $p(x)$ is a pdf that looks like $p(x) = cf(x)$, where $c$ is a constant and $f$ is a continuous function of $x$. Since*

$$\int_x p(x)\,dx = \int_x cf(x)\,dx = 1,$$

*then*

$$\int_x f(x)dx = 1/c.$$

*Note: No calculus is needed to compute $\int_x f(x)\,dx$ if $f(x)$ looks like a known pdf.*

**Example 2.16:** Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes, and this is very rare, since the frequency of occurrence of the gene is very low in human populations.

Consider a woman who has an affected brother (xY), which implies that her mother must be a carrier of the hemophilia gene (xX). We are also told that her father is not affected (XY), thus the woman herself has a fifty-fifty chance of having the gene.

Let $\theta$ denote the state of the woman. It can take two values: the woman is a carrier ($\theta = 1$) or not ($\theta = 0$). Based on this, the prior can be written as

$$P(\theta = 1) = P(\theta = 0) = 1/2.$$

Suppose the woman has a son who does not have hemophilia ($S1 = 0$). Now suppose the woman has another son. Calculate the probability that this second son also will not have hemophilia ($S2 = 0$), given that the first son does not have hemophilia. Assume son one and son two are conditionally independent given $\theta$.

*Solution*:

$$p(S2 = 0|S1 = 0) = \sum_\theta p(S2 = 0|\theta)p(\theta|S1 = 0).$$

First compute

$$p(\theta|S1=0) = \frac{p(S1=0|\theta)p(\theta)}{p(S1=0|\theta=0)p(\theta=0)+p(S1=0|\theta=1)p(\theta=1)}$$
$$= \begin{cases} \frac{(1)(1/2)}{(1)(1/2)+(1/2)(1/2)} = \frac{2}{3} & \text{if } \theta = 0 \\ \frac{1}{3} & \text{if } \theta = 1. \end{cases}$$

Then

$$p(S2=0|S1=0) = p(S2=0|\theta=0)p(\theta=0|S1=0) + p(S2=0|\theta=1)p(\theta=1|S1=0)$$
$$= (1)(2/3) + (1/2)(1/3) = 5/6.$$

**Negative Binomial Distribution**

Before doing the next example, we will introduce the Negative Binomial distribution. The binomial distribution counts the numbers of successes in a fixed number of iid Bernoulli trials. Recall, a Bernoulli trial has a fixed success probability $p$.

Suppose instead that we count the number of Bernoulli trials required to get a fixed number of successes. This formulation leads to the *Negative Binomial distribution.*

In a sequence of independent Bernoulli($p$) trials, let $X$ denote the trial at which the $r$th success occurs, where $r$ is a fixed integer.

Then

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \; x = r, r+1, \ldots$$

and we say $X \sim$ Negative Binom$(r, p)$.

There is another useful formulation of the Negative Binomial distribution. In many cases, it is defined as $Y$ = number of failures before the $r$th success. This formulation is statistically equivalent to the one given above in term of $X$ = trial at which the $r$th success occurs, since $Y = X - r$. Then

$$f(y) = \binom{r+y-1}{y} p^r (1-p)^y, \; y = 0, 1, 2, \ldots$$

and we say $Y \sim$ Negative Binom$(r, p)$.

When we refer to the Negative Binomial distribution in this class, we will refer to the second one defined unless we indicate otherwise.

**Example 2.17:** (Poisson-Gamma)

$$\begin{aligned} X|\lambda &\sim \text{Poisson}(\lambda) \\ \lambda &\sim \text{Gamma}(a, b) \end{aligned}$$

Assume that $\tilde{X}|\lambda \sim \text{Poisson}(\lambda)$ is independent of $X$. Assume we have a new observation $\tilde{x}$. Find the posterior predictive distribution, $p(\tilde{x}|x)$. Assume that $a$ is an integer.

*Solution:*

First, we must find $p(\lambda|x)$.

Recall

$$\begin{aligned} p(\lambda|x) &\propto p(x|\lambda)(p(\lambda) \\ &\propto e^{-\lambda}\lambda^{x}\lambda^{a-1}e^{-\lambda/b} \\ &= \lambda^{x+a-1}e^{-\lambda(1+1/b)}. \end{aligned}$$

Thus, $\lambda|x \sim \text{Gamma}(x + a, \frac{1}{1+1/b})$, i.e., $\lambda|x \sim \text{Gamma}(x + a, \frac{b}{b+1})$.

It then follows that

$$\begin{aligned}
p(\tilde{x}|x) &= \int_\lambda p(\tilde{x}|\lambda)p(\lambda|x)\, d\lambda \\
&= \int_\lambda \frac{e^{-\lambda}\lambda^{\tilde{x}}}{\tilde{x}!}\frac{1}{\Gamma(x+a)(\frac{b}{b+1})^{x+a}}\lambda^{x+a-1}e^{-\lambda(b+1)/b}\, d\lambda \\
&= \frac{1}{\tilde{x}!\,\Gamma(x+a)(\frac{b}{b+1})^{x+a}}\int_\lambda \lambda^{\tilde{x}+x+a-1}e^{-\lambda(2b+1/b)}\, d\lambda \\
&= \frac{1}{\tilde{x}!\,\Gamma(x+a)(\frac{b}{b+1})^{x+a}}\Gamma(\tilde{x}+x+a)(b/(2b+1)^{\tilde{x}+x+a} \\
&= \frac{\Gamma(\tilde{x}+x+a)(b/(2b+1)^{\tilde{x}+x+a}}{\tilde{x}!\,\Gamma(x+a)(\frac{b}{b+1})^{x+a}} \\
&= \frac{\Gamma(\tilde{x}+x+a)}{\tilde{x}!\,\Gamma(x+a)}\,\frac{b^{\tilde{x}+x+a}}{b^{x+a}}\frac{(b+1)^{x+a}}{(2b+1)^{\tilde{x}+x+a}} \\
&= \frac{(\tilde{x}+x+a-1)!}{(x+a-1)!\,\tilde{x}!}\frac{b^{\tilde{x}}\,(b+1)^{x+a}}{(2b+1)^{\tilde{x}+x+a}} \\
&= \binom{\tilde{x}+x+a-1}{\tilde{x}}\left(\frac{b}{2b+1}\right)^{\tilde{x}}\left(\frac{b+1}{2b+1}\right)^{x+a}.
\end{aligned}$$

Let $p = b/(2b+1)$, which implies $1-p = (b+1)/(2b+1)$.

Then

$$p(\tilde{x}|x) = \binom{\tilde{x}+x+a-1}{\tilde{x}}p^{\tilde{x}}(1-p)^{x+a}.$$

Thus,

$$\tilde{x}|x \sim \text{Negative Binom}\left(x+a, \frac{b}{2b+1}\right),$$

where we are assuming the Negative Binomial distribution as defined in Wikipedia (and not as defined earlier in the notes).

**Example 2.18:** Suppose that $X$ is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest modeling it with a Poisson likelihood.

To use a Bayesian analysis, we require a prior distribution for $\theta$ having support on the positive real line. A convenient choice is given by the Gamma distribution, since it's conjugate for the Poisson likelihood.

The model is given by

$$\begin{aligned} X|\lambda &\sim \text{Poisson}(\lambda) \\ \lambda &\sim \text{Gamma}(a, b). \end{aligned}$$

We are also told 42 moms are observed arriving at the particular hospital during December 2007. Using prior study information given, we are told $a = 5$ and $b = 6$. (We found $a, b$ by working backwards from a prior mean of 30 and prior variance of 180).

We would like to find several things in this example:

1. Plot the likelihood, prior, and posterior distributions as functions of $\lambda$ in `R`.
2. Plot the posterior predictive distribution where the number of pregnant women arriving falls between [0,100], integer valued.
3. Find the posterior predictive probability that the number of pregnant women arrive is between 40 and 45 (inclusive).

*Solution*: The first thing we need to know to do this problem are $p(\lambda|x)$ and $p(\tilde{x}|x)$. We found these in Example 2.17. So,

$$\lambda|x \sim \text{Gamma}\left(x + a, \frac{b}{b+1}\right),$$

and

$$\tilde{x}|x \sim \text{Negative Binom}\left(x + a, \frac{b}{2b+1}\right).$$

Next, we can move right into R for our analysis.

```
setwd("~/Desktop/sta4930/ch3")
lam = seq(0,100, length=500)
x = 42
a = 5
b = 6
like = dgamma(lam,x+1,scale=1)
prior = dgamma(lam,5,scale=6)
post = dgamma(lam,x+a,scale=b/(b+1))
pdf("preg.pdf", width = 5, height = 4.5)
plot(lam, post, xlab = expression(lambda), ylab= "Density", lty=2, lwd=3, type="l")
lines(lam,like, lty=1,lwd=3)
lines(lam,prior, lty=3,lwd=3)
legend(70,.06,c("Prior", "Likelihood","Posterior"), lty = c(2,1,3),
lwd=c(3,3,3))
dev.off()

##posterior predictive distribution
xnew = seq(0,100) ## will all be ints
post_pred_values = dnbinom(xnew,x+a,b/(2*b+1))
plot(xnew, post_pred_values, type="h", xlab = "x", ylab="Posterior Predictive Distribution")

## what is posterior predictive prob that number
of pregnant women arrive is between 40 and 45 (inclusive)

(ans = sum(post_pred_values[41:46])) ##recall we included 0
```

In the first part of the code, we plot the posterior, likelihood, and posterior. This should be self-explanatory since we have already done an example.

When we find our posterior predictive distribution, we must create a sequence of integers from 0 to 100 (inclusive) using the `seq` command. Then we find the posterior predictive values using the function `dnbinom`. Then we simply plot the sequence of $x_{\text{new}}$ on the x-axis and the corresponding posterior predictive values on the y-axis. We set `type="h"` so that our plot will appear somewhat like a smooth histogram.

Finally, in order to calculate the posterior predictive probability that the number of pregnant women who arrive is between 40 and 45, we simply add up the posterior predictive probabilities that correspond to these values. We find that the posterior predictive probability of 0.1284 that the number of pregnant women who arrive is between 40 and 45.

# Chapter 3

# Being Objective

*No, it does not make sense for me to be an 'Objective Bayesian'!*
—Stephen E. Fienberg

Thus far in this course, we have mostly considered *informative* or *subjective* priors. Ideally, we want to choose a prior reflecting our beliefs about the unknown parameter of interest. This is a *subjective* choice. All Bayesians agree that wherever prior information is available, one should try to incorporate a prior reflecting this information as much as possible. We have mentioned how incorporation of a prior expert opinion would strengthen purely data-based analysis in real-life decision problems. Using prior information can also be useful in problems of statistical inference when your sample size is small or you have a high- or infinite-dimensional parameter space.

However, in dealing with real-life problems you may run into problems such as

- not having past historical data,
- not having an expert opinion to base your prior knowledge on (perhaps your research is cutting-edge and new), or
- as your model becomes more complicated, it becomes hard to know what priors to put on each unknown parameter.

The problems we have dealt with all semester have been very simple in nature. We have only had one parameter to estimate (except for one example).

Think about a more complex problem such as the following (we looked at this problem in Chapter 1):

$$X|\theta \sim N(\theta, \sigma^2)$$
$$\theta|\sigma^2 \sim N(\mu, \tau^2)$$
$$\sigma^2 \sim \text{IG}(a, b)$$

where now $\theta$ and $\sigma^2$ are both unknown and we must find the posterior distributions of $\theta|X, \sigma^2$ and $\sigma^2|X$. For this slightly more complex problem, it is much harder to think about what values $\mu, \tau^2, a, b$ should take for a particular problem. What should we do in these type of situations?

Often no reliable prior information concerning $\theta$ exists, or inference based completely on the data is desired. It might appear that inference in such settings would be impossible, but reaching this conclusion is too hasty.

Suppose we could find a distribution $p(\theta)$ that contained no or little information about $\theta$ in the sense that it didn't favor one value of $\theta$ over another (provided this is possible). Then it would be natural to refer to such a distribution as a *noninformative prior*. We could also argue that all or most of the information contained in the posterior distribution, $p(\theta|x)$, came from the data. Thus, all resulting inferences were *objective* and not subjective.

Definition 3.1: *Informative/subjective priors* represent our prior beliefs about parameter values before collecting any data. For example, in reality, if statisticians are unsure about specifying the prior, they will turn to the experts in the field or experimenters to look at past data to help fix the prior.

**Example 3.1:** (Pregnant Mothers) Suppose that $X$ is the number of pregnant mothers arriving at a hospital to deliver their babies during a given month. The discrete count nature of the data as well as its natural interpretation leads to adopting a Poisson likelihood,

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \{0, 1, 2, \ldots\},\ \theta > 0.$$

A convenient choice for the prior distribution here is a Gamma$(a, b)$ since it is conjugate for the Poisson likelihood. To illustrate the example further, suppose that 42 moms deliver babies during the month of December. Suppose from past data at this hospital, we assume a prior of Gamma$(5, 6)$. From this, we can easily calculate the posterior distribution, posterior mean and variance, and do various calculations of interest in `R`.

DEFINITION 3.2: *Noninformative/objective priors* contain little or no information about $\theta$ in the sense that they do not favor one value of $\theta$ over another. Therefore, when we calculate the posterior distribution, most if not all of the inference will arise from the likelihood. Inferences in this case are *objective and not subjective*. Let's look at the following example to see why we might consider such priors.

**Example 3.2:** (Pregnant Mothers Continued) Recall Example 3.1. As we noted earlier, it would be natural to take the prior on $\theta$ as Gamma$(a, b)$ since it is the conjugate prior for the Poisson likelihood, however suppose that for this data set we do not have any information on the number of pregnant mothers arriving at the hospital so there is no basis for using a Gamma prior or any other *informative* prior. In this situation, we could take some noninformative prior.

**Comment**: Since many of the objective priors are improper, so we must check that the posterior is proper.

**Theorem 3.1.** *Propriety of the Posterior*

- *If the prior is proper, then the posterior will* always *be proper.*
- *If the prior is improper, you must check that the posterior is proper.*

## ○ Meaning Of Flat

What does a "flat prior" really mean? People really abuse the word flat and interchange it for noninformative. Let's talk about what people really mean when they use the term "flat," since it can have different meanings.

**Example 3.3:** Often statisticians will refer to a prior as being flat, when a plot of its density actually looks flat, i.e., uniform. An example of this would be taking such a prior to be

$$\theta \sim \text{Unif}(0, 1).$$

We can plot the density of this prior to see that the density is flat.

What happens if we consider though the transformation to $1/\theta$. Is our prior still flat?

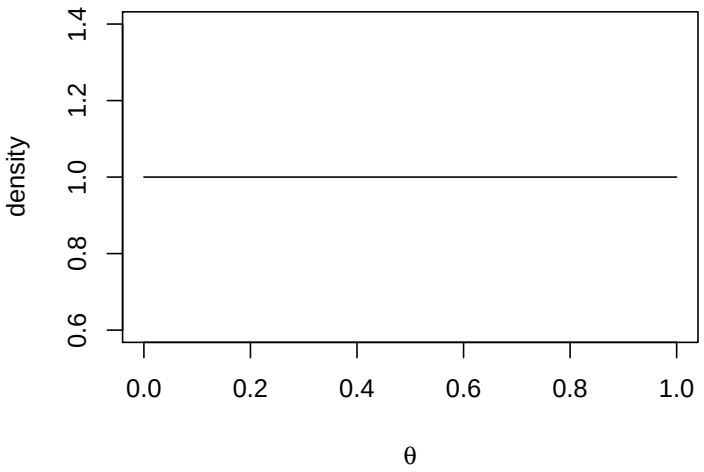



FIGURE 3.1: Unif(0,1) prior

**Example 3.4:** Now suppose we consider Jeffreys' prior, $p_J(\theta)$, where $X \sim \text{Bin}(n, \theta)$.

We calculate Jeffreys' prior by finding the Fisher information. The Fisher information tells us how much information the data gives us for certain parameter values.

In this example, it can be shown that $p_J(\theta) \propto \text{Beta}(1/2, 1/2)$. Let's consider the plot of this prior. Flat here is a purely abstract idea. In order to achieve objective inference, we need to compensate more for values on the boundary than values in the middle.

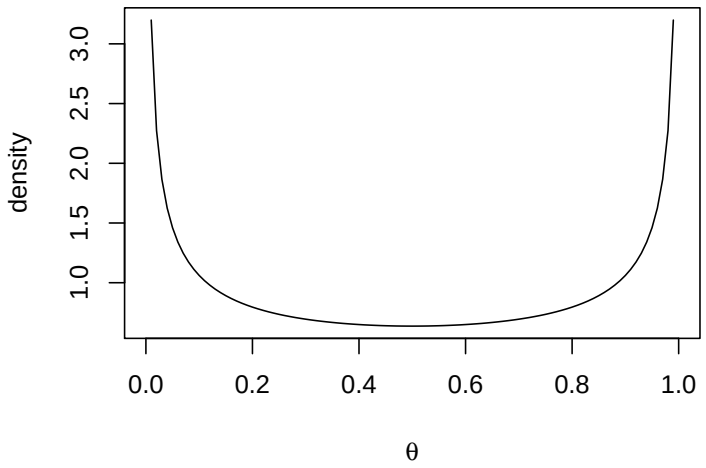



FIGURE 3.2: Jeffreys' prior for Binom likelihood

**Example 3.5:** Finally, we consider the following prior on $\theta$ :

$$\theta \sim N(0, 1000).$$

What happens in this situation? We look at two plots in Figure 3.3 to consider the behavior of this prior.

## ○ Objective Priors in More Detail

### Uniform Prior of Bayes and Laplace

**Example 3.6:** (Thomas Bayes) In 1763, Thomas Bayes considered the question of what prior to use when estimating a binomial success probability $p$. He described the problem quite differently back then by considering throwing balls onto a billiard table. He separated the billiard table into many different intervals and considered different events. By doing so (and not going into the details of this), he argued that a Uniform(0,1) prior was appropriate for $p$.

**Example 3.7:** (Laplace) In 1814, Pierre-Simon Laplace wanted to know

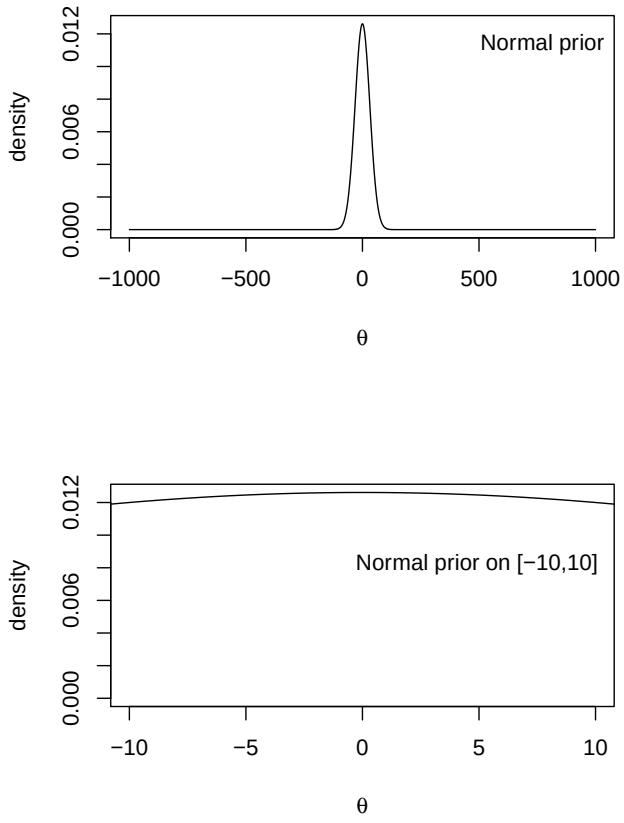



FIGURE 3.3: Normal priors

the probability that the sun will rise tomorrow. He answered this question using the following Bayesian analysis:

- Let $X$ represent the number of days the sun rises. Let $p$ be the probability the sun will rise tomorrow.
- Let $X|p \sim \text{Bin}(n, p)$.
- Suppose $p \sim \text{Uniform}(0, 1)$.
- Based on reading the Bible, Laplace computed the total number of days $n$ in recorded history, and the number of days $x$ on which the sun rose. Clearly, $x = n$.

Then

$$\begin{aligned}\pi(p|x) &\propto \binom{n}{x} p^x (1-p)^{n-x} \cdot 1 \\ &\propto p^{x+1-1}(1-p)^{n-x+1-1}\end{aligned}$$

This implies

$$p|x \sim \text{Beta}(x+1, n-x+1)$$

Then

$$\hat{p} = E[p|x] = \frac{x+1}{x+1+n-x+1} = \frac{x+1}{n+2} = \frac{n+1}{n+2}.$$

Thus, Laplace's estimate for the probability that the sun rises tomorrow is $(n+1)/(n+2)$, where $n$ is the total number of days recorded in history. For instance, if so far we have encountered 100 days in the history of our universe, this would say that the probability the sun will rise tomorrow is $101/102 \approx 0.9902$. However, we know that this calculation is ridiculous. Here, we have extremely strong subjective information (the laws of physics) that says it is extremely likely that the sun will rise tomorrow. Thus, objective Bayesian methods shouldn't be recklessly applied to every problem we study—especially when subjective information this strong is available.

### Criticism of the Uniform Prior

The Uniform prior of Bayes and Laplace and has been criticized for many different reasons. We will discuss one important reason for criticism and not go into the other reasons since they go beyond the scope of this course.

In statistics, it is often a good property when a rule for choosing a prior is *invariant* under what are called one-to-one transformations. Invariant basically means unchanging in some sense. The invariance principle means that a rule for choosing a prior should provide equivalent beliefs even if we consider a transformed version of our parameter, like $p^2$ or $\log p$ instead of $p$.

### Jeffreys' Prior

One prior that is invariant under one-to-one transformations is Jeffreys' prior.

What does the invariance principle mean? Suppose our prior parameter is $\theta$, however we would like to transform to $\phi$.

Define $\phi = f(\theta)$, where $f$ is a one-to-one function.

Jeffreys' prior says that if $\theta$ has the distribution specified by Jeffreys' prior for $\theta$, then $f(\theta)$ will have the distribution specified by Jeffreys' prior for $\phi$. We will clarify by going over two examples to illustrate this idea.

Note, for example, that if $\theta$ has a Uniform prior, Then one can show $\phi = f(\theta)$ will not have a Uniform prior (unless $f$ is the identity function).

Aside from the invariance property of Jeffreys' prior, in the univariate case, Jeffreys' prior satisfies many optimality criteria that statisticians are interested in.

Definition 3.3: Define

$$I(\theta) = -E\left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2}\right],$$

where $I(\theta)$ is called the Fisher information. Then *Jeffreys' prior* is defined to be

$$p_J(\theta) = \sqrt{I(\theta)}.$$

**Example 3.8:** (Uniform Prior is Not Invariant to Transformation)
Let $\theta \sim \text{Uniform}(0, 1)$. Suppose now we would like to transform from $\theta$ to $\theta^2$.

Let $\phi = \theta^2$. Then $\theta = \sqrt{\phi}$. It follows that

$$\frac{\partial \theta}{\partial \phi} = \frac{1}{2\sqrt{\phi}}.$$

Thus, $p(\phi) = \frac{1}{2\sqrt{\phi}}$, $0 < \phi < 1$ which shows that $\phi$ is not Uniform on $(0, 1)$. Hence, the transformation is not invariant. Criticism such as this led to consideration of Jeffreys' prior.

**Example 3.9:** (Jeffreys' Prior Invariance Example)
Suppose

$$X|\theta \sim \text{Exp}(\theta).$$

One can show using calculus that $I(\theta) = 1/\theta^2$. Then $p_J(\theta) = 1/\theta$. Suppose that $\phi = \theta^2$. It follows that

$$\frac{\partial \theta}{\partial \phi} = \frac{1}{2\sqrt{\phi}}.$$

Then

$$\begin{aligned} p_J(\phi) &= p_J(\sqrt{\phi}) \left| \frac{\partial \theta}{\partial \phi} \right| \\ &= \frac{1}{\sqrt{\phi}} \frac{1}{\sqrt{2\phi}} \propto \frac{1}{\phi}. \end{aligned}$$

Hence, we have shown for this example, that Jeffreys' prior is invariant under the transformation $\phi = \theta^2$.

**Example 3.10:** (Jeffreys' prior) Suppose

$$X|\theta \sim \text{Binomial}(n, \theta).$$

Let's calculate the posterior using Jeffreys' prior. To do so we need to calculate $I(\theta)$. Ignoring terms that don't depend on $\theta$, we find

$$\begin{aligned} \log p(x|\theta) &= x \log(\theta) + (n - x) \log(1 - \theta) \implies \\ \frac{\partial \log p(x|\theta)}{\partial \theta} &= \frac{x}{\theta} - \frac{n - x}{1 - \theta} \\ \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} &= -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2} \end{aligned}$$

Since, $E(X) = n\theta$, then

$$I(\theta) = -E\left[-\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta}\frac{n}{(1-\theta)} = \frac{n}{\theta(1-\theta)}.$$

This implies that

$$\begin{aligned} p_J(\theta) &= \sqrt{\frac{n}{\theta(1-\theta)}} \\ &\propto \text{Beta}(1/2, 1/2). \end{aligned}$$

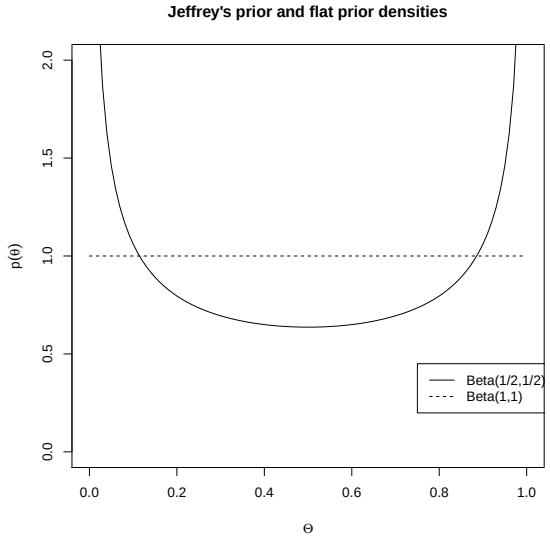



FIGURE 3.4: Jeffreys' prior and flat prior densities

Figure 3.4 compares the prior density $\pi_J(\theta)$ with that for a flat prior, which is equivalent to a Beta(1,1) distribution.

Note that in this case the prior is inversely proportional to the standard deviation. Why does this make sense?

We see that the data has the least effect on the posterior when the true $\theta = 0.5$, and has the greatest effect near the extremes, $\theta = 0$ or 1. Jeffreys' prior compensates for this by placing more mass near the extremes of the range, where the data has the strongest effect. We could get the same effect by (for example) letting the prior be $\pi(\theta) \propto \dfrac{1}{\text{Var}\theta}$ instead of $\pi(\theta) \propto \dfrac{1}{[\text{Var}\theta]^{1/2}}$.

However, the former prior is not invariant under reparameterization, as we would prefer.

We then find that

$$\begin{aligned} p(\theta \mid x) &\propto \theta^x (1-\theta)^{n-x} \theta^{1/2-1} (1-\theta)^{1/2-1} \\ &= \theta^{x-1/2} (1-\theta)^{n-x-1/2} \\ &= \theta^{x-1/2+1-1} (1-\theta)^{n-x-1/2+1-1}. \end{aligned}$$

Thus, $\theta|x \sim \text{Beta}(x + 1/2, n - x + 1/2)$, which is a proper posterior since the prior is proper.

*Note*: Remember that it is important to check that the posterior is proper.

**Jeffreys' and Conjugacy**

Jeffreys priors are widely used in Bayesian analysis. In general, they are not conjugate priors; the fact that we ended up with a conjugate Beta prior for the binomial example above is just a lucky coincidence. For example, with a Gaussian model $X \sim N(\mu, \sigma^2)$, it can be shown that $\pi_J(\mu) = 1$ and $\pi_J(\sigma) = \frac{1}{\sigma}$, which do not look anything like a Gaussian or an inverse gamma, respectively. However, it can be shown that Jeffreys priors are limits of conjugate prior densities. For example, a Gaussian density $N(\mu_o, \sigma_o^2)$ approaches a flat prior as $\sigma_o^2 \to \infty$, while the inverse gamma $\sigma^{-(a+1)} e^{-b/\sigma} \to \sigma^{-1}$ as $a, b \to 0$.

**Limitations of Jeffreys'**

Jeffreys' priors work well for single-parameter models, but not for models with multidimensional parameters. By analogy with the one-dimensional case, one might construct a naive Jeffreys prior as the joint density:

$$\pi_J(\theta) = |I(\theta)|^{1/2},$$

where $|\cdot|$ denotes the determinant and the $(i, j)$th element of the Fisher information matrix is given by

$$I(\theta)_{ij} = -E\left[\frac{\partial^2 \log p(X|\theta)}{\partial \theta_i \partial \theta_j}\right].$$

Let's see what happens when we apply a Jeffreys' prior for $\theta$ to a multivariate Gaussian location model. Suppose $X \sim N_p(\theta, I)$, and we are interested in

performing inference on $||\theta||^2$. In this case the Jeffreys' prior for $\theta$ is flat. It turns out that the posterior has the form of a non-central $\chi^2$ distribution with $p$ degrees of freedom. The posterior mean given one observation of $X$ is $E(||\theta||^2|X) = ||X||^2 + p$. This is not a good estimate because it adds $p$ to the square of the norm of $X$, whereas we might normally want to shrink our estimate towards zero. By contrast, the minimum variance frequentist estimate of $||\theta||^2$ is $||X||^2 - p$.

Intuitively, a multidimensional flat prior carries a lot of information about the expected value of a parameter. Since most of the mass of a flat prior distribution is in a shell at infinite distance, it says that we expect the value of $\theta$ to lie at some extreme distance from the origin, which causes our estimate of the norm to be pushed further away from zero.

### Haldane's Prior

In 1963, Haldane introduced the following improper prior for a binomial proportion:

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

It can be shown to be improper using simple calculus, which we will not go into. However, the posterior is proper under certain conditions. Let

$$Y|\theta \sim \text{Bin}(n, \theta).$$

Calculate $p(\theta|y)$ and show that it is improper when $y = 0$ or $y = n$.

> Remark: Recall that for a Binomial distribution, $Y$ can take values $y = 0, 1, 2, \ldots, n$.

We will first calculate $p(\theta|y)$.

$$\begin{aligned} p(\theta|y) &\propto \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\theta(1-\theta)} \\ &\propto \theta^{y-1}(1-\theta)^{n-y-1} \\ &= \theta^{y-1}(1-\theta)^{(n-y)-1}. \end{aligned}$$

The density of a Beta$(a, b)$ is the following:

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta > 0.$$

This implies that $\theta|Y \sim \text{Beta}(y, n-y)$.

Finally, we need to check that our posterior is proper. Recall that the parameters of the Beta need to be positive. Thus, $y > 0$ and $n - y > 0$. This means that $y \neq 0$ and $y \neq n$ in order for the posterior to be proper.

> Remark: Recall that the Beta density must integrate to 1 whenever the parameter values are positive. Hence, when they are not positive, the density does not integrate to 1 and integrates to $\infty$. Thus, for the problem above, when $y = 0$ and $y = n$ the density is improper.

There are many other objective priors that are used in Bayesian inference, however, this is the level of exposure that we will cover in this course. If you're interested in learning more about objective priors ($g$-prior, probability matching priors), see me and I can give you some references.

## 3.1 Reference Priors

Reference priors were proposed by Jose Bernardo in a 1979 paper, and further developed by Jim Berger and others from the 1980s through the present. They are credited with bringing about an objective Bayesian renaissance; an annual conference is now devoted to the objective Bayesian approach.

The idea behind reference priors is to formalize what exactly we mean by an uninformative prior: it is a function that maximizes some measure of distance or divergence between the posterior and prior, as data observations are made. Any of several possible divergence measures can be chosen, for example the Kullback-Leibler divergence or the Hellinger distance. By maximizing the divergence, we allow the data to have the maximum effect on the posterior estimates.

For one-dimensional parameters, it will turn out that reference priors and Jeffreys' priors are equivalent. For multidimensional parameters, they differ. One might ask, how can we choose a prior to maximize the divergence between the posterior and prior, without having seen the data first? Reference priors handle this by taking the expectation of the divergence, given a model distribution for the data. This sounds superficially like a frequentist approach—basing inference on imagined data. But once the prior is chosen based on some model, inference proceeds in a standard Bayesian fashion.

(This contrasts with the frequentist approach, which continues to deal with imagined data even after seeing the real data!)

### ○ Laplace Approximation

Before deriving reference priors in some detail, we go through the Laplace approximation which is very useful in Bayesian analysis since we often need to evaluate integrals of the form

$$\int g(\theta) f(x|\theta)\pi(\theta)\, d\theta.$$

For example, when $g(\theta) = 1$, the integral reduces to the marginal likelihood of $x$. The posterior mean requires evaluation of two integrals $\int \theta f(x|\theta)\pi(\theta)\, d\theta$ and $\int f(x|\theta)\pi(\theta)\, d\theta$. Laplace's method is a technique for approximating integrals when the integrand has a sharp maximum.

*Remark: There is a nice refinement of the Laplace approximation due to Tierney, Kass, and Kadane (JASA, 1989). Due to time constraints, we won't go into this, but if you're looking to apply this in research, this is something you should look up in the literature and use when needed.*

**Theorem 3.2.** *Laplace Approximation*
*Let $I = \int q(\theta)\exp\{nh(\theta)\}\, d\theta$. Assume that $\hat{\theta}$ maximizes $\theta$ and that $h$ has a sharp maximum at $\hat{\theta}$. Let $c = h''(\hat{\theta}) > 0$. Then*

$$I = q(\hat{\theta})\exp\{nh(\hat{\theta})\}\frac{\sqrt{2\pi}}{\sqrt{nc}}(1 + O(n^{-1})) \approx q(\hat{\theta})\exp\{nh(\hat{\theta})\}\frac{\sqrt{2\pi}}{\sqrt{nc}}$$

*Proof.* Apply Taylor expansion about $\hat{\theta}$.

$$\begin{aligned} I &\approx \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \left[q(\hat{\theta}) + (\theta-\hat{\theta})q'(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 q''(\hat{\theta})\right] \\ &\qquad \times \left[\exp\{nh(\hat{\theta}) + n(\theta-\hat{\theta})h'(\hat{\theta}) + \frac{n}{2}(\theta-\hat{\theta})^2 h''(\hat{\theta})\}\right] d\theta + \cdots \\ &\approx q(\hat{\theta})e^{nh(\hat{\theta})} \int \left[1 + (\theta-\hat{\theta})\frac{q'(\hat{\theta})}{q(\hat{\theta})} + \frac{1}{2}(\theta-\hat{\theta})^2\frac{q''(\hat{\theta})}{q(\hat{\theta})}\right] \\ &\qquad \times \exp\left[\frac{-nc}{2}(\theta-\hat{\theta})^2\}\right] d\theta + \cdots . \end{aligned}$$

Now let $t = \sqrt{nc}(\theta - \hat{\theta})$. This implies that $d\theta = \frac{1}{\sqrt{nc}}dt$. Hence,

$$\begin{aligned} I &\approx \frac{q(\hat{\theta})e^{nh(\hat{\theta})}}{\sqrt{nc}} \int_{-\delta\sqrt{nc}}^{\delta\sqrt{nc}} \left[1 + \frac{t}{\sqrt{nc}}\frac{q'(\hat{\theta})}{q(\hat{\theta})} + \frac{t^2}{2nc}\frac{q''(\hat{\theta})}{q(\hat{\theta})}\right] e^{-t^2/2}\, dt \\ &\approx \frac{q(\hat{\theta})e^{nh(\hat{\theta})}}{\sqrt{nc}}\sqrt{2\pi}\left[1 + 0 + \frac{q''(\hat{\theta})}{q(\hat{\theta})}\frac{1}{2nc}\right] \\ &\approx \frac{q(\hat{\theta})e^{nh(\hat{\theta})}}{\sqrt{nc}}\sqrt{2\pi}\left[1 + O(1/n)\right] \approx \frac{q(\hat{\theta})e^{nh(\hat{\theta})}}{\sqrt{nc}}\sqrt{2\pi}. \end{aligned}$$

□

## ○ Some Probability Theory

First, we give a few definitions from probability theory (you may have seen these before) and we will be informal about these.

- If $X_n$ is $O(n^{-1})$ then $X_n$ "goes to 0 at least as fast as $1/n$."
- If $X_n$ is $o(n^{-1})$ then $X_n$ "goes to 0 faster than $1/n$."

Definition 3.4: Formally, writing

$$X_n = o(r_n) \text{ as } n \to \infty$$

means that

$$\frac{X_n}{r_n} \to 0.$$

Similarly,

$$X_n = O(r_n) \text{ as } n \to \infty$$

means that

$$\frac{X_n}{r_n} \text{ is bounded.}$$

This shouldn't be confused with the definition below:

Definition 3.5: Formally, let $X_n, n \geq 1$ be random vectors and $R_n, n \geq 1$ be positive variables. Then

$$X_n = o_p(R_n) \text{ if } \frac{X_n}{R_n} \xrightarrow{p} 0$$

and

$$X_n = O_p(R_n) \text{ if } \frac{X_n}{R_n} \text{ is bounded in probability.}$$

Recall that $X_n$ is bounded in probability if $\{P_n, n \geq 1\}$ is uniformly tight, where $P_n(A) = Pr(X_n \in A), A \in R^k$, i.e., given any $\epsilon > 0$ there exists an $M$ such that $Pr(\|X_n\| \leq M) \geq 1 - \epsilon$ for all $n \geq 1$. For full details and examples, see Billingsley or van der Vaart.

### ○ Shrinkage Argument of J.K. Ghosh

This argument given by J.K. Ghosh will be used to derive reference priors. It can be used in many other theoretical proofs in Bayesian theory. If interested in seeing these, please refer to his book for details as listed on the syllabus. Please note that below I am hand waving over some of the details regarding analysis that are important but not completely necessary to grasp the basic concept here.

We consider a possibly vector-valued r.v. X with pdf $g(\cdot|\theta)$. Our goal is to find an expression for $E_\theta[q(X, \theta)]$ for some function $q(X, \theta)$, where the integral $\int q(x, \theta)\ f(x|\theta)\ d\theta$ is too difficult to calculate directly. There are three steps to go through to find the desired quantity. The steps are outlined without proof.

**Step 1**: Consider a proper prior $\bar{\pi}(\cdot)$ for $\theta$ such that the support of $\bar{\pi}(\cdot)$ is a compact rectangle in the parameter space and $\bar{\pi}(\cdot)$ vanishes on the boundary of the support, while remaining positive on the interior. Consider the posterior of $\theta$ under $\bar{\pi}(\cdot)$ and hence obtain $E^{\bar{\pi}}[q(X, \theta)|x]$.

**Step 2**: Find $E_\theta E^{\bar{\pi}}[q(x, \theta)|x] = \lambda(\theta)$ for $\theta$ in the interior of the support of $\bar{\pi}(\cdot)$.

**Step 3**: Integrate $\lambda(\cdot)$ with respect to $\bar{\pi}(\cdot)$ and then allow $\bar{\pi}(\cdot)$ to converge to the degenerate prior at the true value of $\theta$ (say $\theta_0$) supposing that the true $\theta$ is an interior point of the support of $\bar{\pi}(\cdot)$. This yields $E_\theta[q(X, \theta)]$.

## ○ Reference Priors

Bernardo (1979) suggested choosing the prior to maximize the expected Kullback-Leibler divergence between the posterior and prior,

$$E\left[\log \frac{\pi(\theta|x)}{\pi(\theta)}\right],$$

where expectation is taken over the joint distribution of $X$ and $\theta$. It is shown in Berger and Bernardo (1989) that if one does this maximization for fixed $n$, it may lead to a discrete prior with finitely many jumps—a far cry from a diffuse prior. Instead, the maximization must be done asymptotically, i.e., as $n \to \infty$. This is achieved as follows:

First write

$$\begin{aligned} E\left[\log \frac{\pi(\theta|x)}{\pi(\theta)}\right] &= \int \int \left[\log \frac{\pi(\theta|x)}{\pi(\theta)}\right] \pi(\theta|x)\, m(x)\, dx\, d\theta \\ &= \int \int \log \frac{\pi(\theta|x)}{\pi(\theta)} f(x|\theta)\, \pi(\theta)\, dx\, d\theta \\ &= \int \pi(\theta) E\left[\log \frac{\pi(\theta|x)}{\pi(\theta)} \,\middle|\, \theta\right] d\theta. \end{aligned}$$

Consider $E\left[\log \frac{\pi(\theta|x)}{\pi(\theta)} \,\middle|\, \theta\right] = E\left[\log \pi(\theta|x) \mid \theta\right] - \log \pi(\theta)$.
Then by iterated expectation,

$$\begin{aligned} E\left[\log \frac{\pi(\theta|x)}{\pi(\theta)}\right] &= \int \pi(\theta) \left\{E\left[\pi(\theta|x); \mid \theta\right] - \log \pi(\theta)\right\}\, d\theta \\ &= \int E\left[\pi(\theta|x)\mid \theta\right] \pi(\theta)\, d\theta - \int \log \pi(\theta)\pi(\theta)\, d\theta. \end{aligned} \tag{3.1}$$

Since we cannot calculate the integral

$$\int E\left[\pi(\theta|x)\mid \theta\right] \pi(\theta)\, d\theta =: q(X, \theta)$$

and we will use Step 1 of the Shrinkage argument of J.K. Ghosh to find $E^{\bar{\pi}}[q(X,\theta)|x]$.

**Step 1**: Find $E^{\bar{\pi}}\left[\log \pi(\theta|x)|x\right] = \int \log \pi(\theta|x)\bar{\pi}(\theta|x)\, d\theta$.

Let $L_n(\theta)$ be defined such that $f(x|\theta) = \exp\{L_n(\theta)\}$.

$$\begin{aligned}\pi(\theta|x) &= \frac{\exp\{L_n(\theta)\}\pi(\theta)}{\int \exp\{L_n(\theta)\}\pi(\theta)\, d\theta} \\ &= \frac{\exp\{L_n(\theta) - L_n(\hat{\theta}_n)\}\pi(\theta)}{\int \exp\{L_n(\theta) - L_n(\hat{\theta}_n)\}\pi(\theta)\, d\theta},\end{aligned}$$

where $\hat{\theta}_n$ denotes the maximum likelihood estimator. Let $t = \sqrt{n}(\theta - \hat{\theta}_n)$, so that $\theta = \hat{\theta}_n + n^{-1/2}t$ and $d\theta = n^{-1/2}\, dt$. We now substitute in for $\theta$ and then perform a Taylor expansion. Recall that in general $f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(x) + \cdots$. Then

$$\begin{aligned}\pi(\theta|x) &= \frac{\exp\{L_n(\hat{\theta}_n + n^{-1/2}t) - L_n(\hat{\theta}_n)\}\pi(\hat{\theta}_n + n^{-1/2}t)}{\int \exp\{L_n(\hat{\theta}_n + n^{-1/2}t) - L_n(\hat{\theta}_n)\}\pi(\hat{\theta}_n + n^{-1/2}t)\, n^{-1/2}\, dt} \\ &= \frac{\exp\{L_n(\hat{\theta}_n) + n^{-1/2}t\, L_n'(\hat{\theta}_n) + n^{-1}t^2\, L_n''(\hat{\theta}_n) - L_n(\hat{\theta}_n) + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)}{\int \exp\{L_n(\hat{\theta}_n) + n^{-1/2}t\, L_n'(\hat{\theta}_n) + n^{-1}t^2\, L_n''(\hat{\theta}_n) - L_n(\hat{\theta}_n) + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)\, n^{-1/2}\, dt} \\ &= \frac{\exp\{n^{-1/2}t\, L_n'(\hat{\theta}_n) + n^{-1}t^2\, L_n''(\hat{\theta}_n) + \cdots\}\pi(\hat{\theta}_n + t/\sqrt{n})}{\int \exp\{n^{-1/2}t\, L_n'(\hat{\theta}_n) + n^{-1}t^2\, L_n''(\hat{\theta}_n) + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)\, n^{-1/2}\, dt}.\end{aligned}$$

Since $\hat{\theta}_n$ is the maximum likelihood estimate, $L_n'(\hat{\theta}_n) = 0$. Now define the quantity

$$\hat{I}_n := \hat{I}_n(\hat{\theta}_n) = -\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\bigg|_{\theta=\hat{\theta}_n} = -L_n''(\hat{\theta}_n).$$

Also, under mild regularity conditions,

$$\hat{I}_n(\hat{\theta}_n)^{1/2}\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I_p).$$

Then we have that

$$\begin{aligned}\pi(\theta|x) &= \frac{\exp\{n^{-1/2}t\, L_n'(\hat{\theta}_n) + \frac{1}{2}t^2\, L_n''(\hat{\theta}_n) + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)}{\int \exp\{n^{-1/2}t\, L_n'(\hat{\theta}_n) + \frac{1}{2}t^2\, L_n''(\hat{\theta}_n) + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)\, n^{-1/2}\, dt}. \\ &= \frac{\exp\{-\frac{1}{2}t^2\hat{I}_n + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)}{\int \exp\{-\frac{1}{2}t^2\hat{I}_n + \cdots\}\pi(\hat{\theta}_n + n^{-1/2}t)\, n^{-1/2}\, dt} \\ &= \frac{\exp\{-\frac{1}{2}t^2\hat{I}_n + O(n^{-1/2})\}[\pi(\hat{\theta}_n) + O(n^{-1/2})]}{\int \exp\{-\frac{1}{2}t^2\hat{I}_n + O(n^{-1/2})\}[\pi(\hat{\theta}_n) + O(n^{-1/2})]\, n^{-1/2}\, dt} \\ &= \frac{\sqrt{n}\exp\{-\frac{1}{2}t^2\hat{I}_n\}\pi(\hat{\theta}_n)[1 + O(n^{-1/2})]}{\sqrt{2\pi}\hat{I}_n^{-1/2}\pi(\hat{\theta}_n)[1 + O(n^{-1/2})]},\end{aligned}$$

noting that the denominator takes the form of a constant times the integral of a normal density with variance $\hat{I}_n^{-1}$. Hence,

$$\pi(\theta|x) = \frac{\sqrt{n}\hat{I}_n^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\hat{I}_n\right)[1 + O(n^{-1/2})]. \tag{3.2}$$

Then $\log \pi(\theta|x) = \frac{1}{2}\log n - \log\sqrt{2\pi} - \frac{1}{2}t^2\hat{I}_n + \frac{1}{2}\log \hat{I}_n + \log[1 + O(n^{-1/2})] = \frac{1}{2}\log n - \log\sqrt{2\pi} - \frac{1}{2}t^2\hat{I}_n + \frac{1}{2}\log \hat{I}_n + \log[O(n^{-1/2})]$. Now consider

$$E^{\bar{\pi}} \log \pi(\theta|x) = \frac{1}{2}\log n - \log\sqrt{2\pi} - E^{\bar{\pi}}\left[\frac{1}{2}t^2\hat{I}_n\right] + \frac{1}{2}\log \hat{I}_n + \log[O(n^{-1/2})].$$

To evaluate $E^{\bar{\pi}}\left[\frac{1}{2}t^2\hat{I}_n\right]$, note that (3.2) states that, up to order $n^{-1/2}$, $\pi(t|x)$ is approximately normal with mean zero and variance $\hat{I}_n^{-1}$. Since this does not depend on the form of the prior $\pi$, it follows that $\bar{\pi}(t|x)$ is also approximately normal with mean zero and variance $\hat{I}_n^{-1}$, again up to order $n^{-1/2}$. Then $E^{\bar{\pi}}\left[\frac{1}{2}t^2\hat{I}_n\right] = \frac{1}{2}$, which implies that

$$\begin{aligned} E^{\bar{\pi}} \log \pi(\theta|x) &= \frac{1}{2}\log n - \log\sqrt{2\pi} - \frac{1}{2} + \log \hat{I}_n^{1/2} + \log[O(n^{-1/2})] \\ &= \frac{1}{2}\log n - \frac{1}{2}\log\sqrt{2\pi e} + \log \hat{I}_n^{1/2} + \log[O(n^{-1/2})]. \end{aligned}$$

**Step 2**: Calculate $\lambda(\theta) = \int E^{\bar{\pi}} \log \pi(\theta|x) f(x|\theta)\, dx$. This is simply

$$\lambda(\theta) = \frac{1}{2}\log n - \frac{1}{2}\log\sqrt{2\pi e} + \log\left[I(\theta)\right]^{1/2} + \log[O(n^{-1/2})].$$

**Step 3**: Since $\lambda(\theta)$ is continuous, the process of calculating $\int \lambda(\theta)\, \bar{\pi}(\theta)\, d\theta$ and allowing $\bar{\pi}(\cdot)$ to converge to degeneracy at $\theta$ simply yields $\lambda(\theta)$ again. Thus,

$$E\left[\pi(\theta|x) \mid \theta\right] = \frac{1}{2}\log n - \frac{1}{2}\log\sqrt{2\pi e} + \log\left[I(\theta)\right]^{1/2} + \log[O(n^{-1/2})].$$

Thus, returning to (3.1), the quantity we need to maximize

$$1/2\log n - \frac{1}{2}\log\sqrt{2\pi e} + \int \log\left\{\frac{\left[I(\theta)\right]^{1/2}}{\pi(\theta)}\right\}\pi(\theta)\, d\theta + \log[O(n^{-1/2})].$$

The integral is non-positive and is maximized above when it is 0, or rather when $\pi(\theta) = I^{1/2}(\theta)$, i.e., Jeffreys' prior.

Take away: If there are no nuisance parameters, Jeffreys' prior is the reference prior.

**Multiparameter generalization**

In the absence of nuisance parameters, the K-L divergence simplifies to

$$E\left[\frac{\pi(\theta|x)}{\pi(\theta)}\right] = \frac{p}{2}\log n - \frac{p}{2}\log(2\pi e) + \int \log\left(\frac{|I(\theta)|^{1/2}}{\pi(\theta)}\right)\pi(\theta)\, d\theta + O(n^{-1/2}).$$

Note that this is maximized when $\pi(\theta) = |I(\theta)|^{1/2}$, meaning that Jeffreys' prior is the maximizer in distance between the prior and posterior. In the presence of nuisance parameters, things change considerably.

**Example 3.11:** Bernardo's reference prior, 1979, *JASA*
Let $\theta = (\theta_1, \theta_2)$, where $\theta_1$ is $p_1 \times 1$ and $\theta_2$ is $p_2 \times 1$. We define $p = p_1 + p_2$. Let

$$I(\theta) = I(\theta_1, \theta_2) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}.$$

Suppose that $\theta_1$ is the parameter of interest and $\theta_2$ is a nuisance parameter (meaning that it's not really of interest to us in the model).

Begin with $\pi(\theta_2|\theta_1) = |I_{22}(\theta)|^{1/2} c(\theta_1)$, where $c(\theta_1)$ is the constant that makes this distribution a proper density. Now try to maximize

$$E\left[\frac{\log \pi(\theta_1|x)}{\pi(\theta_1)}\right]$$

to find the marginal prior $\pi(\theta_1)$. We write

$$\begin{aligned} \log \frac{\pi(\theta_1|x)}{\pi(\theta_1)} &= \log \frac{\pi(\theta_1, \theta_2|x)/\pi(\theta_2|\theta_1, x)}{\pi(\theta_1, \theta_2)/\pi(\theta_2|\theta_1)} \\ &= \log \frac{\pi(\theta|x)}{\pi(\theta)} - \log \frac{\pi(\theta_2|\theta_1, x)}{\pi(\theta_2|\theta_1)}. \end{aligned} \tag{3.3}$$

Arguing as before,

$$E\left[\log \frac{\pi(\theta|x)}{\pi(\theta)}\right] = \frac{p}{2} \log n - \frac{p}{2} \log 2\pi e + \int \pi(\theta) \log \frac{|I(\theta)|^{1/2}}{\pi(\theta)}\, d\theta + O(n^{-1/2}). \tag{3.4}$$

Similarly,

$$E\left[\log \frac{\pi(\theta_2|\theta_1, x)}{\pi(\theta_2|\theta_1)}\right] = \frac{p_2}{2} \log n - \frac{p_2}{2} \log 2\pi e + \int \pi(\theta) \log \frac{|I_{22}(\theta)|^{1/2}}{\pi(\theta_2|\theta_1)}\, d\theta + O(n^{-1/2}). \tag{3.5}$$

From (3.3)–(3.5), we find

$$E\left[\log \frac{\pi(\theta_1|x)}{\pi(\theta_1)}\right] = \frac{p_1}{2} \log n - \frac{p_1}{2} \log 2\pi e + \int \pi(\theta) \log \frac{|I_{11.2}(\theta)|^{1/2}}{\pi(\theta_1)}\, d\theta + O(n^{-1/2}), \tag{3.6}$$

where $I_{11.2}^{1/2}(\theta) = I_{11}(\theta) - I_{12}(\theta) I_{22}^{-1}(\theta) I_{21}(\theta)$ and $I(\theta) = |I_{22}|\, |I_{11} - I_{12} I_{22}^{-1} I_{21}| = |I_{22}|\, |I_{11.2}|$. These can be derived from Searle's book on matrix algebra as a reference.

We now break up the integral in (3.6) and we define

$$\log \psi(\theta_1) = \int \pi(\theta_2|\theta_1) \log |I_{11.2}(\theta)|^{1/2} \, d\theta_2.$$

We find that

$$\begin{aligned} E\left[\log \frac{\pi(\theta_1|x)}{\pi(\theta_1)}\right] &= \frac{p_1}{2}\log n - \frac{p_1}{2}\log 2\pi e + \int \pi(\theta) \log |I_{11.2}(\theta)|^{1/2} \, d\theta \\ &\quad - \int \pi(\theta) \log \pi(\theta_1) \, d\theta + O(n^{-1/2}) \\ &= \frac{p_1}{2}\log n - \frac{p_1}{2}\log 2\pi e + \int \pi(\theta_1) \left[\int \pi(\theta_2|\theta_1) \log |I_{11.2}(\theta)|^{1/2} \, d\theta_2\right] d\theta_1 \\ &\quad - \int \pi(\theta_1) \log \pi(\theta_1) \, d\theta_1 + O(n^{-1/2}) \\ &= \frac{p_1}{2}\log n - \frac{p_1}{2}\log 2\pi e + \int \pi(\theta_1) \log \frac{\psi(\theta_1)}{\pi(\theta_1)} \, d\theta_1 + O(n^{-1/2}). \end{aligned}$$

To maximize the integral above, we choose $\pi(\theta_1) = \psi(\theta_1)$. Note that $I^{-1}_{11.2}(\theta) = I_{11}(\theta)$ where

$$I^{-1}(\theta) = \begin{pmatrix} I_{11}(\theta) & -I_{11}(\theta) I_{12}^{-1}(\theta) I_{22}(\theta) \\ -I_{22}(\theta) I_{21}^{-1}(\theta) I_{11}(\theta) & I_{22}(\theta) \end{pmatrix}.$$

Writing out our prior, we find that

$$\pi(\theta_1) = \exp\{\int \pi(\theta_2|\theta_1) \log |I_{11.2}(\theta)|^{1/2} \, d\theta_2\} = \exp\{\int |I_{22}(\theta)|^{1/2} \log |I_{11.2}(\theta)|^{1/2} \, d\theta_2\}.$$

**Remark**: An important point that should be highlighted is that all these calculations (especially evaluations of all integrals) are carried out over an increasing sequence of compact sets $K$, whose union is the parameter space. For example, if the parameter space is $\mathbb{R} \times \mathbb{R}^+$, take the increasing sequence of compact rectangles $[-i, i] \times [-i^{-1}, i]$ and then eventually take $i \to \infty$. Also, the proofs are carried out by considering a sequence of priors $\pi_i$ with support $K_i$ and we eventually take $i \to \infty$. This fact should be taken into account when doing the examples and calculations of these types of problems.

**Example 3.12:** Let $X_1 \ldots X_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$, where $\sigma^2$ is a nuisance parameter. Consider the sequence of priors $\pi_i$ with support $[-i, i] \times [-i^{-1}, i], i = 1, 2, \ldots$.

$$I(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

Then $\pi(\sigma|\mu) = \frac{\sqrt{2}c_{i2}}{\sigma}$, $i^{-1} \leq \sigma \leq i$. Consider $\int_{i^{-1}}^{i} \frac{\sqrt{2}c_{i2}}{\sigma} = 1 \implies c_{i2} = \frac{1}{2\sqrt{2}\ln i}$. Thus, $\pi(\sigma|\mu) = \frac{1}{2\ln i}\frac{1}{\sigma}$, $i^{-1} \leq \sigma \leq i$. Now find $\pi(\mu)$. Observe that

$$\pi(\mu) = \exp\{\int \pi(\sigma|\mu) \log |I_{11.2}(\theta)|^{1/2}\, d\sigma\}$$

Recall that

$I_{11.2} = I_{11} - I_{12}I_{22}^{-1}I_{21} = 1/\sigma^2$. Thus, $\pi(\mu) = \exp\{\int_{i^{-1}}^{i} c_{i2}\frac{\sqrt{2}}{\sigma}\log(\frac{1}{\sigma})\}\, d\sigma +$ constant $= c$. We want to find $\pi(\mu, \sigma)$. We know that $\pi(\sigma|\mu) = \frac{\pi(\mu,\sigma)}{\pi(\mu)} \implies$ $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma|\mu) = \frac{c}{2\ln i}\frac{1}{\sigma} \propto \frac{1}{\sigma}$.

**Problems with Reference Priors** See page 128 of the little green book. Bernardo and Berger (1992) suggest

## 3.2 Final Thoughts on Being Objective

### Some Thoughts from a Review paper by Stephen Fienberg

We have spent just a short amount of time covering objective Bayesian procedures, but already we have seen how each is flawed in some sense. As Fienberg (2009) points out in a review article (see the webpage), there are two main parts of being Bayesian: the prior and the likelihood. What is Fienberg's point? There is proposed claim that robustness should be carried through at both levels of the model. That is, we should care about subjectivity of the *likelihood as well as the prior.*

What about pragmatism (what works best in terms of implementation)? It really comes down to the quality of the data that we're analyzing. Fienberg looks at two examples. The first involves the NBC Election Night model of the 1960s and 1970s (which used a fully HB model). Here considering different priors is important. For this illustration, there were multiple priors based on past elections to choose from in real time, and the choice of prior was often crucial in close elections. However, in other examples such as Mosteller and Wallace (1964) analysis of the Federalist papers, the likelihood

mattered more. In this case, the posterior odds for several papers shifted when Mosteller and Wallace used a negative binomial versus a Poisson for word counts.

My favorite part that Fienberg illustrates in this paper is the view that "objective Bayes is like the search for the Holy Grail." He mentions that Good (1972) once wrote that there are "46,656 Varieties of Bayesians," which was a number that he admitted exceeded the number of professional statisticians during that time. Today? There seem to be as many choices coming about of objective Bayes for trying to arrive at the perfect choice of an objective prior. Each seems to fail because of foundational principles. For example, Eaton and Freedman (2004) criticize why you shouldn't use Jeffreys' prior for the normal covariance matrix. We didn't look at intrinsic priors, but they have been criticized by Fienberg for contingency tables because of their dependence on the likelihood function and because of bizarre properties when extended to deal with large sparse tables.

Fienberg's conclusion: "No, it does not make sense for me to be an 'Objective Bayesian'!" Read the other papers on the web when you have time and you can make you're own decision.

# Chapter 4

# Evaluating Bayesian Procedures

*They say statistics are for losers, but losers are usually the ones saying that.*
—Urban Meyer

In this chapter, we give a brief overview of how to evaluate Bayesian procedures by looking at how frequentist confidence intervals differ from Bayesian credible intervals. We also introduce Bayesian hypothesis testing and Bayesian p-values. Again, we emphasize that this is a rough overview and for more details, one should look in Gelman et al. (2004) or Carlin and Louis (2009).

## 4.1 Confidence Intervals versus Credible Intervals

One major difference between Bayesians and frequentists is how they interpret intervals. Let's quickly review what a frequentist confidence interval is and how to interpret one.

**Frequentist Confidence Intervals**

A confidence interval for an unknown (fixed) parameter $\theta$ is an interval of numbers that we believe is likely to contain the true value of $\theta$. Intervals are important because they provide us with an idea of how well we can estimate $\theta$.

DEFINITION 4.1: A *confidence interval* is constructed to contain $\theta$ a percentage of the time, say 95%. Suppose our confidence level is 95% and our interval is $(L, U)$. Then we are 95% confident that the true value of $\theta$ is contained in $(L, U)$ in *the long run*. In the long run means that this would occur nearly 95% of the time if we repeated our study millions and millions of times.

**Common Misconceptions in Statistical Inference**

- A confidence interval is a statement about $\theta$ (a population parameter). It is not a statement about the sample.
- Remember that a confidence interval is *not* a statement about individual subjects in the population. As an example, suppose that I tell you that a 95% confidence interval for the average amount of television watched by Americans is (2.69, 6.04) hours. This *doesn't* mean we can say that 95% of all Americans watch between 2.69 and 6.04 hours of television. We also *cannot* say that 95% of Americans in the sample watch between 2.69 and 6.04 hours of television. Beware that statements such as these are false. However, we can say that we are 95 percent confident that the *average* amount of televison watched by Americans is between 2.69 and 6.04 hours.

**Bayesian Credible Intervals**

Recall that frequentists treat $\theta$ as fixed, but Bayesians treat $\theta$ as a random variable. The main difference between frequentist confidence intervals and Bayesian credible intervals is the following:

- Frequentists invoke the concept of probability before observing the data. For any fixed value of $\theta$, a frequentist confidence interval will contain the true parameter $\theta$ with some probability, e.g., 0.95.
- Bayesians invoke the concept of probability after observing the data. For some particular set of data $X = x$, the random variable $\theta$ lies in a Bayesian credible interval with some probability, e.g., 0.95.

### Assumptions

In lower-level classes, you wrote down assumptions whenever you did confidence intervals. This is redundant for any problem we construct in this course since we always know the data is randomly distributed and we assume it comes from some underlying distribution, say Normal, Gamma, etc. We also always assume our observations are i.i.d. (independent and identically distributed), meaning that the observations are all independent and they all have the same variance. Thus, when working a particular problem, we will assume these assumptions are satisfied given the proposed model holds.

Definition 4.2: A Bayesian credible interval of size $1 - \alpha$ is an interval $(a, b)$ such that

$$P(a \leq \theta \leq b|x) = 1 - \alpha.$$

$$\int_a^b p(\theta|x)\, d\theta = 1 - \alpha.$$

Remark: When you're calculating credible intervals, you'll find the values of $a$ and $b$ by several means. You could be asked do the following:

- Find the $a, b$ using means of calculus to determine the credible interval or set.
- Use a Z-table when appropriate.
- Use `R` to approximate the values of $a$ and $b$.
- You could be given `R` code/output and asked to find the values of $a$ and $b$.

**Important Point**

Our definition for the credible interval could lead to many choices of $(a, b)$ for particular problems.

Suppose that we required our credible interval to have equal probability $\alpha/2$ in each tail. That is, we will assume

$$P(\theta < a|x) = \alpha/2$$

and

$$P(\theta > b|x) = \alpha/2.$$

Is the credible interval still unique? No. Consider

$$\pi(\theta|x) = I(0 < \theta < 0.025) + I(1 < \theta < 1.95) + I(3 < \theta < 3.025)$$

so that the density has three separate plateaus. Now notice that any $(a, b)$ such that $0.025 < a < 1$ and $1.95 < b < 3$ satisfies the proposed definition of a ostensibly "unique" credible interval. To fix this, we can simply require that

$$\{\theta : \pi(\theta|x) \text{ is positive}\}$$

(i.e., the support of the posterior) must be an interval.

Bayesian interval estimates for $\theta$ are similar to confidence intervals of classical inference. They are called credible intervals or sets. Bayesian credible intervals have a nice interpretation as we will soon see.

To see this more clearly, see Figure 4.1.

Definition 4.3: A Bayesian credible set $C$ of level or size $1 - \alpha$ is a set $C$ such that $1 - \alpha \leq P(C|y) = \int_C p(\theta|y)\, d\theta$. (Of course in discrete settings, the integral is simply replaced by summation).

Note: We use $\leq$ instead of $=$ to include discrete settings since obtaining exact coverage in a discrete setting may not be possible.

This definition enables direct probability statements about the likelihood of $\theta$ falling in $C$. That is,

"The probability that $\theta$ lies in $C$ given the observed data y is at least $(1 - \alpha)$."

This greatly contrasts with the usual frequentist CI, for which the corresponding statement is something like "If we could recompute $C$ for a large

number of datasets collected in the same way as ours, about $(1-\alpha)\times 100\%$ of them would contain the true value $\theta$. "

This classical statement is not one of comfort. We may not be able to repeat our experiment a large number of times (suppose we have an an interval estimate for the 1993 U.S. unemployment rate). If we are in physical possession of just one dataset, our computed $C$ will either contain $\theta$ or it won't, so the actual coverage probability will be 0 or 1. For the frequentist, the confidence level $(1-\alpha)$ is only a "tag" that indicates the quality of the procedure. But for a Bayesian, the credible set provides an actual probability statement based only on the observed data and whatever prior information we add.

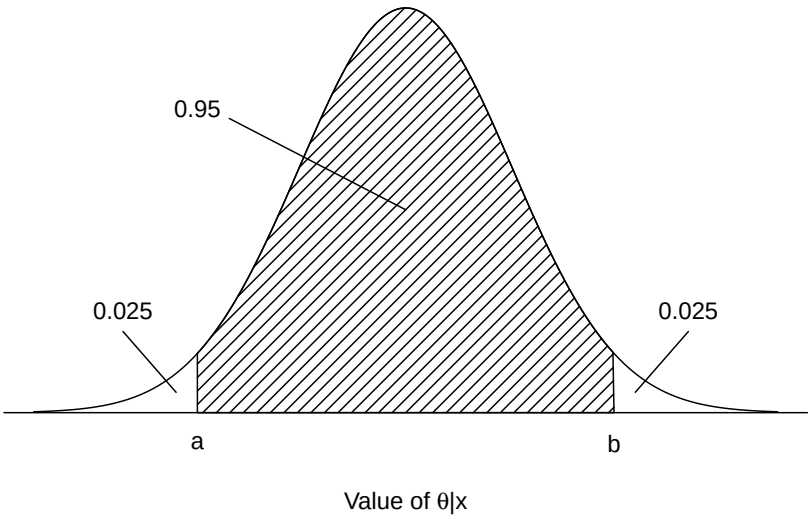



FIGURE 4.1: Illustration of 95% credible interval

**Interpretation**

We interpret Bayesian credible intervals as follows: There is a 95% probability that the true value of $\theta$ is in the interval $(a, b)$, given the data.

**Comparisons**

- Conceptually, probability comes into play in a frequentist confidence interval *before* collecting the data, i.e., there is a 95% probability that we will collect data that produces an interval that contains the true parameter value. However, this is awkward, because we would like to make statements about the probability that the interval contains the true parameter value given the data that we actually observed.
- Meanwhile, probability comes into play in a Bayesian credible interval *after* collecting the data, i.e., based on the data, we now think there is a 95% probability that the true parameter value is in the interval. This is more natural because we want to make a probability statement regarding that data after we have observed it.

**Example 4.1:** Suppose

$$\begin{aligned} X_1 \ldots, X_n|\theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu, \tau^2), \end{aligned}$$

where $\mu, \sigma^2$, and $\tau^2$ are known. Calculate a 95% credible interval for $\theta$.

Recall

$$\theta|x_1, \ldots x_n \sim N\left(\frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\right).$$

Let

$$\begin{aligned} \mu^* &= \frac{n\bar{x}\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}, \\ \sigma^{*2} &= \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}. \end{aligned}$$

We want to calculate $a$ and $b$ such that $P(\theta < a|x_1, \ldots, x_n) = 0.05/2 = 0.025$ and $P(\theta > b|x_1, \ldots, x_n) = 0.05/2 = 0.025$. So,

$$\begin{aligned} 0.025 &= P(\theta < a|x_1, \ldots, x_n) \\ &= P\left(\frac{\theta - \mu^*}{\sigma^*} < \frac{a - \mu^*}{\sigma^*}\middle| x_1, \ldots, x_n\right) \\ &= P\left(Z < \frac{a - \mu^*}{\sigma^*}\middle| x_1, \ldots, x_n\right), \text{ where } Z \sim N(0, 1). \end{aligned}$$

Thus, we now must find an $a$ such that $P\left(Z < \frac{a-\mu^*}{\sigma^*} \middle| x_1, \ldots, x_n\right) = 0.025$. From a Z-table, we know that

$$\frac{a-\mu^*}{\sigma^*} = -1.96.$$

This tells us that $a = \mu^* - 1.96\sigma^*$. Similarly, $b = \mu^* + 1.96\sigma^*$. (Work this part out on your own at home). Therefore, a 95% credible interval is

$$\mu^* \pm 1.96\sigma^*.$$

**Example 4.2:** We're interested in knowing the true average number of ornaments on a Christmas tree. Call this $\theta$. We take a random sample of $n$ Christmas trees, count the ornaments on each one, and call the results $X_1, \ldots, X_n$. Let the prior on $\theta$ be Normal$(75, 225)$.

Using data (`trees.txt`) we have, we will calculate the 95% credible interval and confidence interval for $\theta$. In `R` we first read in the data file `trees.txt`. We then set the initial values for our known parameters, $n, \sigma, \mu$, and $\tau$.

Next, we refer to Example 4.1, and calculate the values of $\mu^*$ and $\sigma^*$ using this example. Finally, again referring to Example 4.1, we recall that the formula for a 95% credible interval here is

$$\mu^* \pm 1.96\sigma^*.$$

On the other hand, recalling back to any basic statistics course, a 95% confidence interval in this situation is

$$\bar{x} \pm 1.96\sigma/\sqrt{n}.$$

From the `R` code, we find that there is a 95% probability that the average number of ornaments per tree is in (45.00, 57.13) given the data. We also find that we are 95% confident that the average number of ornaments per tree is contained in (43.80, 56.20). If we compare the width of each interval, we see that the credible interval is slightly narrower. It is also shifted towards slightly higher values than the confidence interval for this data, which makes sense because the prior mean was higher than the sample mean. What would happen to the width of the intervals if we increased $n$? Does this make sense?

```
x = read.table("trees.txt",header=T)
attach(x)
```

```
n = 10
sigma = 10
mu = 75
tau = 15

mu.star = (n*mean(orn)*tau^2+mu*sigma^2)/(n*tau^2+sigma^2)
sigma.star = sqrt((sigma^2*tau^2)/(n*tau^2+sigma^2))

(cred.i = mu.star+c(-1,1)*qnorm(0.975)*sigma.star)
(conf.i = mean(orn)+c(-1,1)*qnorm(0.975)*sigma/sqrt(n))

diff(cred.i)
diff(conf.i)
detach(x)
```

**Example 4.3:** (Sleep Example)
Recall the Beta-Binomial. Consider that we were interested in the proportion of the population of American college students that sleep at least eight hours each night ($\theta$).

Suppose a random sample of 27 students from UF, where 11 students recorded they slept at least eight hours each night. So, we assume the data is distributed as Binomial$(27, \theta)$.

Suppose that the prior on $\theta$ was Beta(3.3,7.2). Thus, the posterior distribution is

$$\theta|11 \sim \text{Beta}(11 + 3.3, 27 - 11 + 7.2), \text{ i.e.,}$$
$$\theta|11 \sim \text{Beta}(14.3, 23.2).$$

Suppose now we would like to find a 90% credible interval for $\theta$. We cannot compute this in closed form since computing probabilities for Beta distributions involves messy integrals that we do not know how to compute. However, we can use `R` to find the interval.

We need to solve

$$P(\theta < c|x) = 0.05$$

and

$$P(\theta > d|x) = 0.05 \text{ for } c \text{ and } d.$$

The reason we cannot compute this in closed form is because we need to compute

$$\int_0^c \text{Beta}(14.3, 23.2)\, d\theta = 0.05$$

and

$$\int_d^1 \text{Beta}(14.3, 23.2)\, d\theta = 0.05.$$

Note that Beta(14.3,23.2) represents

$$f(\theta) = \frac{\Gamma(37.5)}{\Gamma(14.3)\Gamma(23.2)}\theta^{14.3-1}(1-\theta)^{23.2-1}.$$

The `R` code for this is very straightforward:

```
a = 3.3
b = 7.2
n = 27
x = 11
a.star = x+a
b.star = n-x+b

c = qbeta(0.05,a.star,b.star)
d = qbeta(1-0.05,a.star,b.star)
```

Running the code in `R`, we find that a 90% credible interval for $\theta$ is (0.256, 0.514), meaning that there is a 90% probability that the proportion of UF students who sleep eight or more hours per night is between 0.256 and 0.514 given the data.

## 4.2 Credible Sets or Intervals

Definition 4.4: Suppose the posterior density for $\theta$ is unimodal. A highest posterior density (HPD) credible set of size $1-\alpha$ is a set $C$ such that $C = \{\theta : p(\theta|Y = y) \geq k_\alpha\}$ where $k_\alpha$ is chosen so that $P(\theta \in C) \geq 1-\alpha$.

**Example 4.4:** Normal HPD credible interval
Suppose that

$$y|\theta \sim N(\theta, \sigma), \quad \theta \sim N(\theta, \tau^2).$$

Then $\theta|y \sim N(\mu_1, \tau_1^2)$, where we have derived these before. The HPD credible interval for $\theta$ is simply $\mu \pm z_{\alpha/2}\tau_1$. Note that the HPD credible interval is that same as the equal tailed interval interval centered at the posterior mean.

**Remark**: Credible intervals are very easy to calculate unlike confidence intervals, which require pivotal quantities or inversion of a family of tests.

In general, plot the posterior distribution and find the HPD credible set. One important point is that the posterior must be unimodal in order to guarantee that the HPD credible set is an interval. (Unimodality of the posterior is a sufficient condition for the credible set to be an interval, but it's not necessary.)

**Example 4.5:** Suppose

$$\begin{aligned} y_1, \ldots, y_n|\sigma^2 &\overset{iid}{\sim} N(0, \sigma^2) \\ p(\sigma^2) &\propto (\sigma^2)^{\alpha/2-1} e^{-\frac{\beta}{2\sigma^2}}. \end{aligned}$$

Let $z = 1/\sigma^2$. Then

$$p(z) \propto z^{\alpha/2+1} e^{-\frac{\beta}{2z}} \left|\frac{1}{z^2}\right| = z^{\alpha/2-1} e^{-\frac{\beta}{2z}}.$$

Then

$$p(\sigma^2|y) \propto (\sigma^2)^{-(n+\alpha)/2-1} e^{-\frac{1}{2\sigma^2}(\sum_i y_i^2+\beta)},$$

which implies that

$$\sigma^2|y \sim IG\left((n+\alpha)/2, (\sum_i y_i^2 + \beta)/2\right).$$

This posterior distribution is unimodal, but how do we know this? One way of showing the posterior is unimodal is to show that it is increasing in $\sigma^2$ up to a point and then decreasing afterwards. The log of the posterior has the same feature.

Then

$$\log(p(\sigma^2|y)) = c_1 - [(n+\alpha)/2 + 1]\log(\sigma^2) - \frac{1}{2\sigma^2}(\sum_i y_i^2 + \beta).$$

This implies that

$$\begin{aligned}\frac{\partial \log(p(\sigma^2|y))}{\partial \sigma^2} &= -\frac{n+\alpha+2}{2\sigma^2} + \frac{1}{\sigma^4}(\sum_i y_i^2 + \beta) \\ &= \frac{(\sum_i y_i^2 + \beta) - (n+\alpha+2)\sigma^2}{2\sigma^4},\end{aligned}$$

which is increasing with respect to $\sigma^2$, equal to, or decreasing as $\sigma^2 \geq (\sum_i y_i^2 + \beta)/(n+\alpha+2)$, etc. Thus, the posterior is unimodal, so we can get a HPD interval for $\sigma^2$.

## 4.3 Bayesian Hypothesis Testing

Let's first review p-values and why they might not make sense in the grand scheme of things. In classical statistics, the traditional approach proposed by Fisher, Neyman, and Pearson is where we have a null hypothesis and an alternative. After determining some test statistic $T(y)$, we compute the p-value, which is

$$\text{p-value } = P\left\{T(Y) \text{ is more extreme than } T(y_{\text{obs}}) \mid H_o\right\},$$

where extremeness is in the direction of the alternative hypothesis. If the p-value is less than some pre specified Type I error rate, we reject $H_o$, and otherwise we don't.

Clearly, classical statistics has deep roots and a long history. It's popular with practitioners, but does it make sense? The approach can be applied in a straightforward manner only when the two hypothesis in question are nested (meaning one within the other). This means that $H_o$ must be a simplification of $H_a$. Many practical testing problems involve a choice between two or more models that aren't nested (choosing between quadratic and exponential growth models for example).

Another difficulty is that tests of this type can only offer evidence against the null hypothesis. A small p-value indicates that the later, alternative model has significantly more explanatory power. But a large p-value does not suggest that the two models are equivalent (only that we lack evidence that they are not). This limitation/difficulty is often swept under the rug and never dealt with. We simply say, "we fail to reject the null hypothesis" and leave it at that.

Third, the p-value offers no direct interpretation as a "weight of evidence" but only as a long-term probability of obtaining data at least as unusual as what we observe. Unfortunately, the fact that small p-values imply rejection of $H_o$ causes many consumers of statistical analyses to assume that the p-value is the probability that $H_o$ is true, even though it's nothing of the sort.

Finally, one last criticism is that p-values depend not only on the observed data but also on the total sampling probability of certain unobserved data points, namely, the more extreme $T(Y)$ values. Because of this, two experiments with identical likelihoods could result in different p-values if the two experiments were designed differently. (This violates the Likelihood Principle.) See Example 1.1 in Chapter 1 for an illustration of how this can happen.

In classical settings, we talk about Type I and Type II errors. In Bayesian hypothesis testing, we will consider the following scenarios:

$$H_o : \theta \in \Theta_o \qquad H_a : \theta \in \Theta_1.$$

$$H_o : \theta = \theta_o \qquad H_a : \theta \neq \theta_0.$$

$$H_o : \theta \leq \theta_o \qquad H_a : \theta > \theta_0.$$

A Bayesian talks about posterior odds and Bayes factors.

Definition 4.5: Prior odds
Let $\pi_o = P(\theta \in \Theta_o)$, $\pi_1 = P(\theta \in \Theta_1)$ and $\pi_o + \pi_1 = 1$. Then the prior odds in favor of $H_o = \dfrac{\pi_o}{\pi_1}$.

Definition 4.6: Posterior odds
Let $\alpha_o = P(\theta \in \Theta_o|y)$ and $\alpha_1 = P(\theta \in \Theta_1|y)$. Then the posterior odds $= \dfrac{\alpha_o}{\alpha_1}$.

Definition 4.7: Bayes Factor
The Bayes Factor (BF)$= \dfrac{\text{posterior odds}}{\text{prior odds}} = \dfrac{\alpha_o}{\alpha_1} \div \dfrac{\pi_o}{\pi_1} = \dfrac{\alpha_o \pi_1}{\alpha_1 \pi_o}$.

**Example 4.6:** IQ Scores
Suppose we're studying IQ scores and so we assume that the data follow the model where

$$\begin{aligned} y|\theta &\sim N(\theta, 10^2) \\ \theta &\sim N(100, 15^2). \end{aligned}$$

We'd like to be able to say something about the mean of the IQ scores and whether it's below or larger than 100. Then

$$H_o : \theta \leq 100 \qquad H_a : \theta > 100.$$

The prior odds are then $\frac{\pi_o}{\pi_1} = \frac{P(\theta \leq 100)}{P(\theta > 100)} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$ by symmetry.

Suppose we find that $y = 115$. Then $\theta | y = 115 \sim N(110.39, 63.23)$.

Then $\alpha_o = P(\theta_o \leq 100 | y = 115) = 0.106$ and $\alpha_1 = P(\theta_1 > 100 | y = 115) = 0.894$. Thus, $\frac{\alpha_o}{\alpha_1} = 0.1185$. Hence, BF $= 0.1185$.

### ○ Lavine and Schervish (*The American Statistician*, 1999): Bayes Factors: What They Are and What They Are Not

We present an example from the paper above to illustrate an important point regarding Bayes Factors. Suppose a coin is known to be a 2-sided head, a 2-sided tail, or fair. Then let $\theta$ be the probability of a head $\in \{0, 1/2, 1\}$. Suppose the data tell us that the coin was tossed 4 times and *always* landed on heads.

Furthermore, suppose that

$$\pi(\{0\}) = 0.01,\ \pi(\{1/2\}) = 0.98,\ \pi(\{1\}) = 0.01.$$

Consider

$$H_1 : \theta = 1 \text{ versus } H_4 : \theta \neq 1$$
$$H_2 : \theta = 1/2 \text{ versus } H_5 : \theta \neq 1/2$$
$$H_3 : \theta = 0 \text{ versus } H_6 : \theta \neq 0.$$

Then

$$f(x|H_1) = P(\text{four heads}|\theta = 1) = 1$$
$$f(x|H_2) = P(\text{four heads}|\theta = 1/2) = (\frac{1}{2})^4 = \frac{1}{16}$$
$$f(x|H_3) = P(\text{four heads}|\theta = 0) = 0$$
$$f(x|H_4) = P(\text{four heads}|\theta \neq 1) = \frac{\frac{1}{16} \times 0.98 + 0 \times 0.01}{0.98 + 0.01} = \frac{1}{16} \times \frac{98}{99} = 0.0619$$
$$f(x|H_5) = P(\text{four heads}|\theta \neq 1/2) = \frac{1 \times 0.01 + 0 \times 0.01}{0.98 + 0.01} = 0.05$$
$$f(x|H_6) = P(\text{four heads}|\theta \neq 0) = \frac{\frac{1}{16} \times 0.98 + 1 \times 0.01}{0.98 + 0.01} = 0.072$$

We then find that

$$\frac{f(x|H_1)}{f(x|H_4)} = 1/0.0619 \text{ and } \frac{f(x|H_2)}{f(x|H_5)} = \frac{\frac{1}{16}}{\frac{1}{2}} = 0.125.$$

Let $k \in (0.0619, 0.125)$ and reject if $BF_{01} < k$. Then we reject $H_4$ in favor of $H_1$. We fail to reject $H_2$. Thus, failing to reject $H_2$ implies failing to reject $H_4$. In this example, evidence in favor of $H_4$ should be stronger than that of $H_2$. But the Bayes Factor violates this. Lavine and Schervish refer to this as *lack of coherence.* The problem does not occur with the posterior odds since if

$$P(\Theta_o|x) < P(\Theta_a|x)$$

holds, then

$$\frac{P(\Theta_o|x)}{1 - P(\Theta_o|x)} < \frac{P(\Theta_a|x)}{1 - P(\Theta_a|x)}.$$

(This result can be generalized).

- Bayes factors are insensitive to the choice of prior, however, this statement is misleading. (Berger, 1995) We will see why in Example 4.5.
- BF measures the change from priors odds to the posterior odds.

**Example 4.7:** Simple Null versus Simple Alternative

$$H_o : \theta = \theta_o \qquad H_a : \theta = \theta_1.$$

Then $\pi_o = P(\theta = \theta_o)$ and $\pi_1 = P(\theta = \theta_1)$, so $\pi_o + \pi_1 = 1$.

Then

$$\alpha_o = P(\theta = \theta_o|y) = \frac{P(y|\theta = \theta_o)P(\theta = \theta_o)}{P(y|\theta = \theta_o)P(\theta = \theta_o) + P(y|\theta = \theta_1)P(\theta = \theta_1)} = \frac{P(y|\theta = \theta_o)\pi_o}{P(y|\theta = \theta_o)\pi_o + P(y|\theta = \theta_1)\pi_1}.$$

This implies that $\frac{\alpha_o}{\alpha_1} = \frac{\pi_o\, P(y|\theta = \theta_o)}{\pi_1\, P(y|\theta = \theta_1)}$ and hence BF $= \frac{P(y|\theta = \theta_o)}{P(y|\theta = \theta_1)}$, which is the likelihood ratio. This does not depend on the choice of the prior. However, in general the Bayes factor depends on how the prior spreads mass over the null and alternative (so Berger's statement is misleading).

**Example 4.8:**

$$H_o : \theta \in \theta_o \qquad H_a : \theta \in \theta_1.$$

Derive the BF. Let $g_o(\theta)$ and $g_1(\theta)$ be probability density functions such that $\int_{\Theta_o} g_o(\theta)\, d\theta = 1$ and $\int_{\Theta_1} g_1(\theta)\, d\theta = 1$. Let

$$\pi(\theta) = \begin{cases} \pi_o\, g_o(\theta) & \text{if } \theta \in \Theta_o \\ \pi_1\, g_1(\theta) & \text{if } \theta \in \Theta_1. \end{cases}$$

Then $\int \pi(\theta)\, d\theta = \int_{\Theta_o} \pi_o\, g_o(\theta)\, d\theta + \int_{\Theta_1} \pi_1\, g_1(\theta)\, d\theta = \pi_0 + \pi_1 = 1$.

This implies that $\frac{\alpha_o}{\alpha_1} = \frac{\int_{\Theta_o} \pi(\theta|y)\, d\theta}{\int_{\Theta_1} \pi(\theta|y)\, d\theta}$. Thus, $\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{m(y)}$. This implies that

$$\frac{\alpha_o}{\alpha_1} = \frac{\frac{\int_{\Theta_o} p(y|\theta)\pi(\theta)\, d\theta}{m(y)}}{\frac{\int_{\Theta_1} p(y|\theta)\pi(\theta)\, d\theta}{m(y)}} = \frac{\int_{\Theta_o} p(y|\theta)\pi_o g_o(\theta)\, d\theta}{\int_{\Theta_1} p(y|\theta)\pi_1 g_1(\theta)\, d\theta}$$

$$= \frac{\pi_o \int_{\Theta_o} p(y|\theta) g_o(\theta)\, d\theta}{\pi_1 \int_{\Theta_1} p(y|\theta) g_1(\theta)\, d\theta} \implies$$

BF $= \frac{\int_{\Theta_o} p(y|\theta) g_o(\theta)\, d\theta}{\int_{\Theta_1} p(y|\theta) g_1(\theta)\, d\theta}$, which is the marginal of $y$ under $H_o$ divided by the marginal of $y$ under $H_1$.

## 4.4 Bayesian p-values

Bayes factors are meant to compare two or more models, however, often we are interested in the goodness of fit of a particular model rather than

comparison of the models. Bayesian p-values were proposed to address these problems.

### ○ Prior Predictive p-value

George Box proposed a prior predictive p-value. Suppose that $T(x)$ is a test statistic and $\pi$ is some prior. Then we calculate the marginal distribution

$$m[T(x) \geq T(x_{\text{obs}})|M_o],$$

where $M_o$ is the null model under consideration.

**Example 4.9:** $X_1, \ldots, X_n|\theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$. Let $M_o : \theta = 0$ and $T(x) = \sqrt{n}|\bar{X}|$.

Suppose the prior $\pi(\sigma^2)$ is degenerate at $\sigma_o^2$. Then $\pi(\sigma^2 = \sigma_o^2) = 1$. Marginally,

$$\bar{X} \sim N(0, \sigma^2/n)$$

under $M_o$. Also,

$$P(\sqrt{n}|\bar{X}| \geq \sqrt{n}|\bar{x}_{\text{obs}}|) = P(\frac{\sqrt{n}|\bar{X}|}{\sigma_o} \geq \frac{\sqrt{n}|\bar{x}_{\text{obs}}|}{\sigma_o}) = 2\Phi(-\frac{\sqrt{n}|\bar{x}_{\text{obs}}|}{\sigma_o}).$$

If the guessed $\sigma_o$ is much smaller than the actual model variance, then the p-value is small and the evidence again $M_o$ is overestimated.

**Remark**: The takeaway message is that the prior predictive p-value is heavily influenced by the prior.

### ○ Other Bayesian p-values

Since then, the posterior predictive p-value (PPP) has been proposed by Rubin (1984), Meng (1994), and Gelman et al. (1996). They propose looking at the posterior predictive distribution of a future observation $x$ under some prior $\pi$. That is, we calculate

$$m(x|x_{\text{obs}}) = \int f(x|\theta)\pi(\theta|x_{\text{obs}})\, d\theta.$$

Then PPP is defined to be

$$P^* = P(T(X) \geq T(x_{\text{obs}})),$$

which is the conditional probability that for a future observation $T(X) \geq T(x_{\text{obs}})$ given the predictive distribution of $X$ under prior $\pi$ and $x_{\text{obs}}$.

**Remark**: For details, see the papers. A general criticism by Bayarri and Berger points out that the procedure involves using the data twice. The data is used in finding the posterior distribution of $\theta$ and also in finding the posterior predictive p-value. As an alternative, they have suggested using conditional predictive p-values (CPP). This involves splitting the data into two parts, say $T(X)$ and $U(X)$. We use $U(X)$ to find the posterior predictive distribution and $T(X)$ continues to be the test statistic.

**Potential Fixes to the Prior Predictive p-value**

We first consider the Conditional and Partial Predictive p-values by Bayarri and Berger, JASA, (1999, 2000). They propose splitting the data into two parts $(T(X), U(X)$, where $T$ is the test statistic and the p-value is computed from the posterior predictive distribution of a future $T$ conditional on $U$. The choice of $U$ is unclear and for complex problems it is nearly impossible to find. We note that if $U(X)$ is taken to be the entire data and $T(X)$ is some test statistic, then we get the PPP back.

Also, Robins, van der Waart, and Ventura, JASA, 2000 investigate Bayarri and Berger's claims that for a parametric model, that their conditional and partial predictive p-values are superior to the parametric bootstrap p-value and to previously proposed p-values (prior predictive p-value of Guttman, 1967 and Rubin, 1984 and the discrepancy p-value of Gelman et. al (1995, 1996) and Meng (1994). Robins et. al note that Bayarri and Berger's claims of superiority is based on small-sample properties for specific examples. They investigate large sample properties and conclude that asymptotic results confirm the superiority of the conditional predictive p-value and partial posterior predictive p-values.

Robins et. al (2000) also explore corrections for when these p-values are difficult to compute. In Section 4 of their paper, they discuss how to modify the test statistic for the parametric bootstrap p-value, posterior predictive p-values, and discrepancy p-values. Modifications are made such they are asymptotically uniform. They claim that their approach is successful for the

discrepancy p-value (and the authors derive a test based on this). Note: the discrepancy p-value can be difficult to calculate for complex models.

## 4.5 Appendix to Chapter 4 (Done by Rafael Stern)

**Added Example for Chapter 4 on March 21, 2013**
The following example is an adaptation from Carlos Alberto de Braganc a Pereira (2006).

Consider that $U_1, U_2, U_3, U_4$ are conditionally i.i.d. given $\theta$ and such that $U_1|\theta$ has Uniform distribution on $(\theta - 0.5, \theta + 0.5)$. Next, we construct a confidence interval and a credible interval for $\theta$.

Let's start with a confidence interval. Let $U_{(1)} = \min\{U_1, U_2, U_3, U_4\}$ and $U_{(4)} = \max\{U_1, U_2, U_3, U_4\}$. Let's prove that $(U_{(1)}, U_{(4)})$ is a 87.5% confidence interval for $\theta$.

Consider

$$\begin{aligned} P(\theta \notin (U_{(1)}, U_{(4)})|\theta) &= P(U_{(1)} > \theta \cup U_{(4)} < \theta|\theta) = \\ &= P(U_{(1)} > \theta|\theta) + P(U_{(4)} < \theta|\theta) = \\ &= P(U_i > \theta, i = 1,2,3,4|\theta) + P(U_i < \theta, i = 1,2,3,4|\theta) = \\ &= (0.5)^4 + (0.5)^4 = (0.5)^3 = 0.125 \end{aligned}$$

Hence, $P(\theta \in (U_{(1)}, U_{(4)})|\theta) = 0.875$, which proves that $(U_{(1)}, U_{(4)})$ is a 87.5% confidence interval for $\theta$.

Consider that $U_{(1)} = 0.1$ and that $U_{(4)} = 0.9$. The 87.5% probability has to do with the random interval $(U_{(1)}, U_{(4)})$ and not with the particular observed value of $(0.1, 0.9)$.

Let's do some investigative work! Observe that, for every $u_i$, $u_i > \theta - 0.5$. Hence, $u_{(1)} > \theta - 0.5$, that is, $\theta < u_{(1)} + 0.5$. Similarly, $\theta > u_{(4)} - 0.5$. Hence, $\theta \in (u_{(4)} - 0.5, u_{(1)} + 0.5)$. Plugging in $u_{(1)} = 0.1$ and $u_{(4)} = 0.9$, obtain $\theta \in (0.4, 0.6)$. That is, even though the observed 87.5% confidence interval is $(0.1, 0.9)$, we know that $\theta \in (0.4, 0.6)$ with certainty.

Let's now compute a 87.5% centered credible interval. This depends on the prior for $\theta$. Consider the improper prior $p(\theta) = 1$, $\theta \in \mathbb{R}$. Observe that:

$$\begin{aligned} P(\theta|u_1,u_2,u_3,u_4) &\propto P(\theta)P(u_1,u_2,u_3,u_4|\theta) = \\ &= \prod_{i=1}^{4} I(u_i)_{(\theta-0.5,\theta+0.5)} = \\ &= I(\theta)_{(u_{(4)}-0.5,u_{(1)}+0.5)}. \end{aligned}$$

That is, $\theta|u$ has Uniform distribution on $(u_{(4)} - 0.5, u_{(1)} + 0.5)$. Let $a = u_{(4)} - 0.5$ and $b = u_{(1)} + 0.5$. The centered 87.5% credible interval is $(l, u)$ such that $\int_a^l \frac{1}{b-a}dx = 2^{-4}$ and $\int_u^b \frac{1}{b-a}dx = 2^{-4}$. Hence, $l = a + \frac{b-a}{2^4}$ and $u = b - \frac{b-a}{2^4}$. Observe that this interval is always a subset of $(a, b)$, which we know contains $\theta$ for sure.

Does $\left(U_{(4)} - 0.5 + \frac{1-(U_{(4)}-U_{(1)})}{2^4}, U_{(1)} + 0.5 - \frac{1-(U_{(4)}-U_{(1)})}{2^4}\right)$ have any confidence guarantees? Before getting into troublesome calculations, we can check this through simulations.

The following `R` code generates a barplot for how often the credible interval captures the correct parameter given different parameter values.

```
sim_capture_theta <- function(theta,nsim) {
  samples <- runif(4*nsim,theta-0.5,theta+0.5)
  dim(samples) <- c(nsim,4)

  success <- function(sample)
  {
    aa <- min(sample)
    bb <- max(sample)
    return((theta > aa + (bb-aa)/16) && (theta < bb - (bb-aa)/16))
  }

  return(mean(apply(samples,1,success)))
}

capture_frequency <- sapply(0:1000, function(ii){sim_capture_theta(ii,1000)})
barplot(capture_frequency,
main="Coverage of credible interval for parameter in 0,1,...,1000")
abline(h=0.825,lty=22)
abline(h=0.85,lty=22)
abline(h=0.875,lty=22)
```

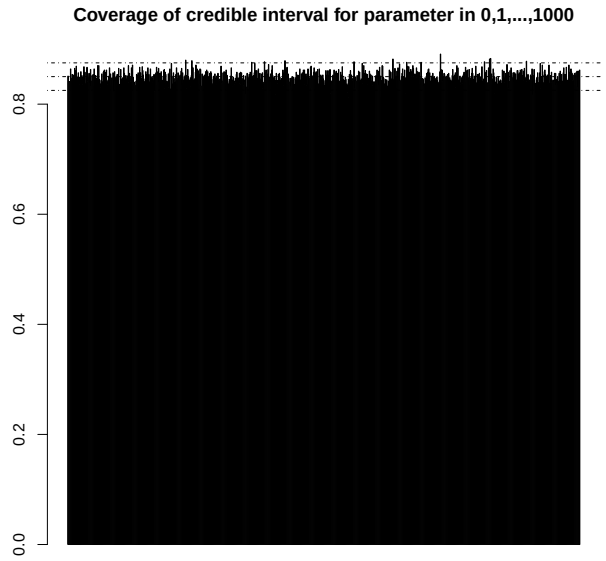



FIGURE 4.2

The result in Figure 4.2 shows that, in this case, the coverage of the credible interval seems to be uniform on the parameter space. This is not guaranteed to always happen! Also, although we constructed a 87.5% credible interval, the picture suggests that $\left(U_{(4)} - 0.5 + \frac{1-(U_{(4)}-U_{(1)})}{2^4}, U_{(1)} + 0.5 - \frac{1-(U_{(4)}-U_{(1)})}{2^4}\right)$ is somewhere near a 85% confidence interval.

Observe that, the wider the gap between $U_{(1)}$ and $U_{(4)}$ the smaller is the region in which $\theta$ can lie. In this sense, it would be nice if the interval would be smaller, the larger this gap is. The example shows that there exist both credible and confidence intevals with this property, but this property isn't achieved by guaranteeing confidence alone.

# Chapter 5

# Monte Carlo Methods

*Every time I think I know what's going on, suddenly there's another layer of complicitions. I just want this damned thing solved.*
—John Scalzi, The Lost Colony

## 5.1 A Quick Review of Monte Carlo Methods

One motivation for Monte Carlo methods is to approximate an integral of the form $\int_X h(x)f(x)\,dx$ that is intractable, where $f$ is a probability density. You might wonder why we wouldn't just use numerical integration techniques. There are a few reasons:

- The most serious problem is the so-called "curse of dimensionality." Suppose we have a $p$-dimensional integral. Numerical integration typically entails evaluating the integrand over some grid of points. However, if $p$ is even moderately large, then any reasonably fine grid will contain an impractically large number of points. For example if $p = 6$, then a grid with just ten points in each dimension—already too coarse for any sensible amount of precision—will consist of $10^6$ points. If $p = 50$, then even an absurdly coarse grid with just *two* points in each dimension will consist of $2^{50}$ points (note that $2^{50} > 10^{15}$).
- There can still be problems even when the dimensionality is small. There are packages in `R` called **area** and **integrate**, however, area

cannot deal with infinite bounds in the integral, and even though integrate can handle infinite bounds, it is fragile and often produces output that's not trustworthy (Robert and Casella, 2010).

### ○ Classical Monte Carlo Integration

The generic problem here is to evaluate $E_f[h(x)] = \int_X h(x)f(x)\,dx$. The classical way to solve this is to generate a sample $(X_1, \ldots, X_n)$ from $f$ and propose as an approximation the empirical average

$$\bar{h}_n = \frac{1}{n}\sum_{j=1}^{n} h(x_j).$$

Why? It can be shown that $\bar{h}_n$ converges a.s. (i.e. for almost every generated sequence) to $E_f[h(X)]$ by the Strong Law of Large Numbers.

Also, under certain assumptions (which we won't get into, see Casella and Robert, page 65, for details), the asymptotic variance can be approximated and then can be estimated from the sample $(X_1, \ldots, X_n)$ by

$$v_n = 1/n^2 \sum_{j=1}^{n} [h(x_j) - \bar{h}_n]^2.$$

Finally, by the CLT (for large $n$),

$$\frac{\bar{h}_n - E_f[h(X)]}{\sqrt{v}_n} \overset{\text{approx.}}{\sim} N(0,1).$$

There are examples in Casella and Robert (2010) along with `R` code for those that haven't seen these methods before or want to review them.

### ○ Importance Sampling

Importance sampling involves generating random variables from a different distribution and then reweighing the output. It's name is given since the new distribution is chosen to give greater mass to regions where $h$ is large (the important part of the space).

Let $g$ be an arbitrary density function and then we can write

$$I = E_f[h(x)] = \int_X h(x)\frac{f(x)}{g(x)}g(x)\,dx = E_g\left[\frac{h(x)f(x)}{g(x)}\right]. \tag{5.1}$$

This is estimated by

$$\hat{I} = \frac{1}{n}\sum_{j=1}^{n}\frac{f(X_j)}{g(X_j)}h(X_j) \longrightarrow E_f[h(X)] \tag{5.2}$$

based on a sample generated from $g$ (not $f$). Since (5.1) can be written as an expectation under $g$, (5.2) converges to (5.1) for the same reason the Monte carlo estimator $\bar{h}_n$ converges.

Remark: Calculating the variance of $\hat{I}$, we find

$$Var(\hat{I}) = \frac{1}{n^2}\sum_i Var\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right) = \frac{1}{n}Var\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right) \Longrightarrow$$
$$\widehat{Var}(\hat{I}) = \frac{1}{n}\widehat{Var}\left(\frac{h(X_i)f(X_i)}{g(X_i)}\right).$$

**Example 5.1:** Suppose we want to estimate $P(X > 5)$, where $X \sim N(0,1)$.

Naive method: Generate n iid standard normals and use the proportion $\hat{p}$ that are larger than 5.

Importance sampling: We will sample from a distribution that gives high probability to the "important region" (the set $(5,\infty)$) and then reweight.

Solution: Let $\phi_o$ and $\phi_\theta$ be the densities of the $N(0,1)$ and $N(\theta,1)$ distributions ($\theta$ taken around 5 will work). We have

$$p = \int I(u>5)\phi_o(u)\,du = \int\left[I(u>5)\frac{\phi_o(u)}{\phi_\theta(u)}\right]\phi_\theta(u)\,du.$$

In other words, if

$$h(u) = I(u>5)\frac{\phi_o(u)}{\phi_\theta(u)}$$

then $p = E_{\phi_\theta}[h(X)]$. If $X_1, \ldots, X_n \sim N(\theta, 1)$, then an unbiased estimate is $\hat{p} = \frac{1}{n}\sum_i h(X_i)$.

We implement this in `R` as follows:

```
1 - pnorm(5)                    # gives 2.866516e-07

# Naive method
set.seed(1)
ss <- 100000
x <- rnorm(n=ss)
phat <- sum(x>5)/length(x)
sdphat <- sqrt(phat*(1-phat)/length(x)) # gives 0

# IS method

set.seed(1)
y <- rnorm(n=ss, mean=5)
h <- dnorm(y, mean=0)/dnorm(y, mean=5) * I(y>5)
mean(h)                         # gives 2.865596e-07
sd(h)/sqrt(length(h))           # gives 2.157211e-09
```

**Example 5.2:** Let $f(x)$ be the pdf of a $N(0,1)$. Assume we want to compute

$$a = \int_{-1}^{1} f(x)dx = \int_{-1}^{1} N(0,1)dx$$

We can use importance sampling to do this calculation. Let $g(X)$ be an arbitrary pdf,

$$a(x) = \int_{-1}^{1} \frac{f(x)}{g(x)} g(x)\, dx.$$

We want to be able to draw $g(x) \sim Y$ easily. But how should we go about choosing $g(x)$?

- Note that if $g \sim Y$, then $a = E[I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}]$.
- The variance of $I_{[-1,1]}(Y)\dfrac{f(Y)}{g(Y)}$ is minimized picking $g \propto I_{[-1,1]}(x)f(x)$. Nevertheless simulating from this $g$ is usually expensive.

- Some $g$'s which are easy to simulate from are the pdf's of the Uniform$(-1, 1)$, the Normal$(0, 1)$ and a Cauchy with location parameter 0.
- Below, there is code of how to get a sample from $I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$ for these distributions,

```
uniformIS <- function(nn) {
  sapply(runif(nn,-1,1),
    function(xx) dnorm(xx,0,1)/dunif(xx,-1,1)) }

cauchyIS <- function(nn) {
  sapply(rt(nn,1),
    function(xx) (xx <= 1)*(xx >= -1)*dnorm(xx,0,1)/dt(xx,2)) }

gaussianIS <- function(nn) {
  sapply(rnorm(nn,0,1),
    function(xx) (xx <= 1)*(xx >= -1)) }
```

Figure 5.1 presents histograms for a sample size 1000 from each of these distributions. The sample variance of $I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$ was, respectively, 0.009, 0.349 and 0.227 (for the uniform, cauchy, and the normal).

- Even though the shape of the uniform distribution is very different from $f(x)$, a standard normal, in $(-1, 1)$, $f(x)$ has a lot of mass outside of $(-1, 1)$.
- This is why the histograms for the Cauchy and the Normal have big bars on 0 and the variance obtained from the uniform distribution is the lowest.
- How would these results change if we wanted to compute the integral over the range $(-3, 3)$ instead of $(-1, 1)$? This is left as a homework exercise.

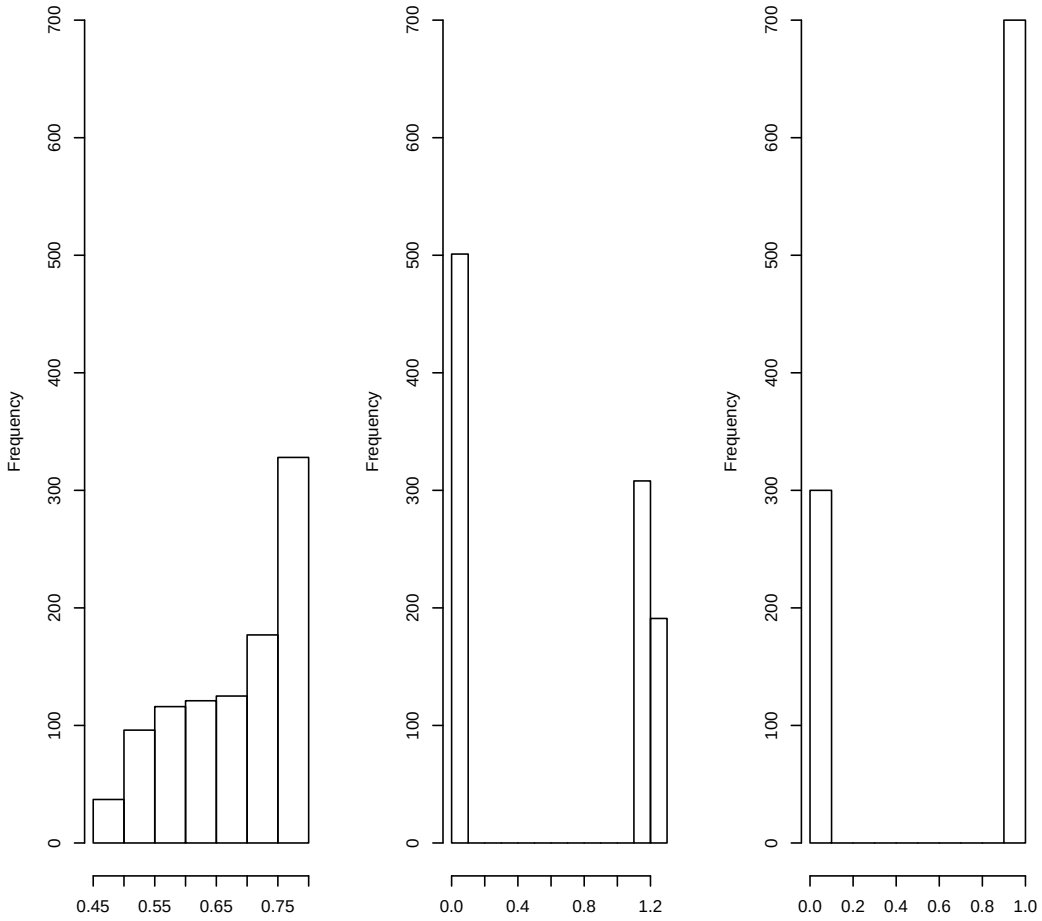



FIGURE 5.1: Histograms for samples from $I_{[-1,1]}(Y)\frac{f(Y)}{g(Y)}$ when $g$ is, respectivelly, a uniform, a Cauchy and a Normal pdf.

### ○ Importance Sampling with unknown normalizing constant

Often we have sample from $\mu$, but know $\pi(x)$ except for a multiplicative $\mu(x)$ constant. Typical example is Bayesian situation:

- $\pi = \nu_Y$ = posterior density of $\theta$ given $Y$ when prior density is $\nu$.
- $\mu = \lambda_Y$ = posterior density of $\theta$ given $Y$ when prior density is $\lambda$.

  We want to estimate $\dfrac{\pi(x)}{\mu(x)} = \dfrac{c_\nu L(\theta)\nu(\theta)}{c_\lambda L(\theta)\lambda(\theta)} = c\dfrac{\nu(\theta)}{\lambda(\theta)} = c\,\ell(x)$,

  where $\ell(x)$ is known and $c$ is unknown.

  Remark: get a ratio of priors.

  Then if we're estimating $h(x)$, we find

$$\begin{aligned}\int h(x)\pi(x)\,dx &= \int h(x)\,c\,\ell(x)\mu(x)\,d(x)\\ &= \frac{\int h(x)\,c\,\ell(x)\mu(x)\,d(x)}{\int \mu(x)\,d(x)}\\ &= \frac{\int h(x)\,c\,\ell(x)\mu(x)\,d(x)}{\int c\,\ell(x)\mu(x)\,d(x)}\\ &= \frac{\int h(x)\,\ell(x)\mu(x)\,d(x)}{\int \ell(x)\mu(x)\,d(x)}.\end{aligned}$$

  Generate $X_1, \ldots, X_n \sim \mu$ and estimate via

$$\frac{\sum_i h(X_i)\,\ell(X_i)}{\sum_i \ell(X_i)} = \sum_i h(X_i)\left(\frac{\ell(X_i)}{\sum_j \ell(X_j)}\right) = \sum_i w_i h(X_i)$$

  where $w_i = \dfrac{\ell(X_i)}{\sum_j \ell(X_j)} = \dfrac{\nu(\theta_i)/\lambda(\theta_i)}{\sum_j \nu(\theta_j)/\lambda(\theta_j)}$.

  **Motivation**
  Why the choice above for $\ell(X)$? Just taking a ratio of priors. The motivation is the following for example:

  – Suppose our application is to Bayesian statistics where $\theta_1, \ldots, \theta_n \sim \lambda_Y$.

- Think about the posterior corresponding here is an essay to deal with conjugate prior $\lambda$.
- Think of $\pi = \nu$ as a complicated prior and $\mu = \lambda$ as a conjugate prior.
- Then the weights are $w_i = \dfrac{\nu(\theta_i)/\lambda(\theta_i)}{\sum_j \nu(\theta_j)/\lambda(\theta_j)}$.

1. If $\mu$ and $\pi$ i.e. $\nu$ and $\lambda$ differ greatly most of the weight will be taken up by a few observations resulting in an unstable estimate.
2. We can get an estimate of the variance of $\dfrac{\sum_i h(X_i)\,\ell(X_i)}{\sum_i \ell(X_i)}$ but we need to use theorems from advance probability theory (The Cramer-Wold device and the Multivariate Delta Method). We'll skip these details.
3. In the application of Bayesian statistics, the cancellation of a potentially very complicated likelihood can lead to a great simplification.
4. The original purpose of importance sampling was to sample more heavily from regions that are important. So, we may do importance sampling using a density $\mu$ because it's more convenient than using a density $\pi$. (These could also be measures if the densities don't exist for those taking measure theory).

### ○ Rejection Sampling

Suppose $\pi$ is a density on the reals and suppose $\pi(x) = c\ l(x)$ where $l$ is known, $c$ is not known. We are interested in case where $\pi$ is complicated. Want to generate $X \sim \pi$.

Motivating idea: look at a very simple case of rejection sampling.

Suppose first that $l$ is bounded and is zero outside of $[0,1]$. Suppose also $l$ is constant on the intervals $((j-1)/k, j/k), j = 1, \ldots, k$. Let $M$ be such that $M \geq l(x)$ for all $x$.

For very simple case, consider the following procedure.

1. Generate a point $(U_1, U_2)$ uniformly at random from the rectangle of height $M$ sitting on top of the interval $[0,1]$.

2. If the point is below the graph of the function $l$, retain $U_1$. Else, reject the point and go back to (1).

Remark: Using the Probability Integral Transformation in reverse. If $X \sim F^{-1}(U)$, then $X \sim F$ where $U \sim \text{Uniform}(0,1)$.

Remark: Think about what this is doing, we're generating many draws that are wasting time. Think about the restriction on $[0,1]$ and if this makes sense.

General Case:

Suppose the density $g$ is such that for some known constant $M$, $Mg(x) \geq l(x)$ for all $x$. Procedure:

1. Generate $X \sim g$, and calculate $r(X) = \dfrac{l(X)}{M\, g(X)}$.
2. Flip a coin with probability of success $r(X)$. If we have a success, retain X. Else return to (1).

To show that an accepted point has distribution $\pi$, let $I$ = indicator that the point is accepted. Then

$$P(I = 1) = \int P(I = 1 \mid X = x) g(x)\, dx = \int \frac{\pi(x)/c}{M\, g(x)} g(x)\, dx = \frac{1}{c\, M}.$$

Thus, if $g_l$ is the conditional distribution of $X$ given $I$, we have

$$g_I(x|I = 1) = g(x) \frac{\pi(x)/c}{M\, g(x)} / P(I = 1) = \pi(x).$$

**Example 5.3:** Suppose we want to generate random variables from the Beta(5.5,5.5) distribution. Note: There are no direct methods for generating from Beta(a,b) if a,b are not integers.

One possibility is to use a Uniform(0,1) as the trial distribution. A better idea is to use an approximating normal distribution.

```
##simple rejection sampler for Beta(5.5,5.5), 3.26.13
```

```
a <- 5.5; b <- 5.5
m <- a/(a+b); s <- sqrt((a/(a+b))*(b/(a+b))/(a+b+1))
funct1 <- function(x) {dnorm(x, mean=m, sd=s)}
funct2 <- function(x) {dbeta(x, shape1=a, shape2=b)}

##plotting normal and beta densities
pdf(file = "beta1.pdf", height = 4.5, width = 5)
plot(funct1, from=0, to=1, col="blue", ylab="")
plot(funct2, from=0, to=1, col="red", add=T)
dev.off()

##M=1.3 (this is trial and error to get a good M)
funct1 <- function(x) {1.3*dnorm(x, mean=m, sd=s)}
funct2 <- function(x) {dbeta(x, shape1=a, shape2=b)}
pdf(file = "beta2.pdf", height = 4.5, width = 5)
plot(funct1, from=0, to=1, col="blue", ylab="")
plot(funct2, from=0, to=1, col="red", add=T)
dev.off()

##Doing accept-reject
##substance of code
set.seed(1); nsim <- 1e5
x <- rnorm(n=nsim, mean=m, sd=s)
u <- runif(n=nsim)
ratio <- dbeta(x, shape1=a, shape2=b) /
           (1.3*dnorm(x, mean=m, sd=s))
ind <- I(u < ratio)
betas <- x[ind==1]
# as a check to make sure we have enough
length(betas) # gives 76836

funct2 <- function(x) {dbeta(x, shape1=a, shape2=b)}
pdf(file = "beta3.pdf", height = 4.5, width = 5)
plot(density(betas))
plot(funct2, from=0, to=1, col="red", lty=2, add=T)
dev.off()
```

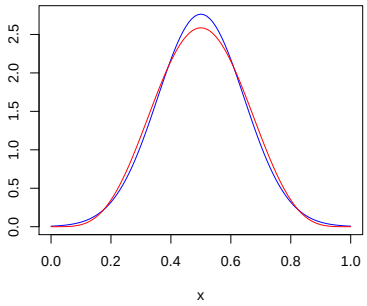



FIGURE 5.2: Normal enveloping Beta

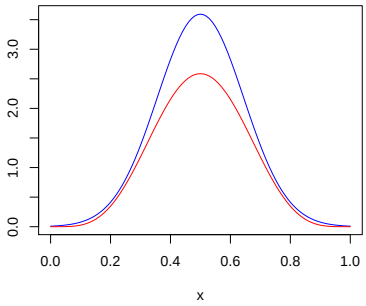



FIGURE 5.3: Naive rejection sampling, M=1.3

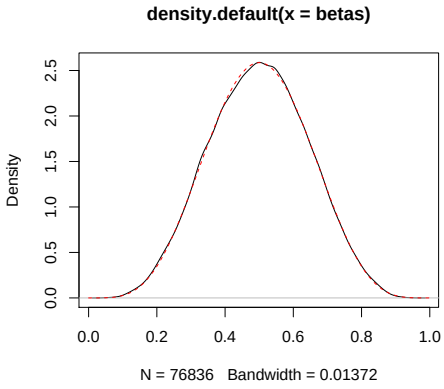



FIGURE 5.4: Rejection sampler

## 5.2 Introduction to Gibbs and MCMC

The main idea here involves iterative simulation. We sample values on a random variable from a sequence of distributions that converge as iterations continue to a target distribution. The simulated values are generated by a Markov chain whose stationary distribution is the target distribution, i.e., the posterior distribution.

Geman and Geman (1994) introduced Gibbs sampling for simulating a multivariate probability distribution $p(x)$ using as random walk on a vector $x$, where $p(x)$ is not necessarily a posterior density.

### ○ Markov Chains and Gibbs Samplers

We have a probability distribution $\pi$ on some space $X$ and we are interested in estimating $\pi$ or $\int h(x)\pi(x)dx$, where h is some function. We are considering situation where $\pi$ is analytically intractable.

**The Basic idea of MCMC**

- Construct a sequence of random variables $X_1, X_2, \ldots$ with the property that the distribution of $X_n$ converges to $\pi$ as $n \to \infty$.
- If $n_o$ is large, then $X_{n_o}, X_{n_0+1} \ldots$ all have the distribution $\pi$ and these can be used to estimate $\pi$ and $\int h(x)\pi(x)dx$.

Two problems:

1. The distribution of $X_{n_o}, X_{n_0+1} \ldots$ is only approximately $\pi$.
2. The random variables $X_{n_o}, X_{n_0+1} \ldots$ are NOT independent; they may be correlated.

**The MCMC Method** Setup: We have a probability distribution $\pi$ which is analytically intractable. Want to estimate $\pi$ or $\int h(x)\pi(x)dx$, where h is some function.

The MCMC method consider of coming up with a transition probability function $P(x, A)$ with the property that it has station distribution $\pi$.

A Markov chain with Markov transition function $P(\cdot, \cdot)$ is a sequence of random variables $X_1, X_2, \ldots$ on a measurable space such that:

1. $P(X_{n+1} \in A | X_n = x) = P(x, A)$.
2. $P(X_{n+1} \in A | X_1, X_2, \ldots, X_n) = P(X_{n+1} \in A | X_n = x)$.

1.) is called a Markov transition function and 2.) is the Markov property, which says "where I'm going next only depends on where I am right now."

Coming back to the MCMC method, we fix a starting point $x_o$ and generate an observation from $X_1$ from $P(x_o, \cdot)$, generate an observation from $X_2$ from $P(X_1, \cdot)$, etc. This generates the Markov chain $x_o = X_o, X_1, X_2, \ldots,$

If we can show that

$$\sup_{C \in B} |P^n(x, C) - \pi(C)| \to 0 \text{ for all } x \in X$$

then by running the chain sufficiently long enough, we succeed in generating an observation $X_n$ with distribution approximately $\pi$.

**What is a Markov chain?**
Start with a sequence of dependent random variables, $\{X^{(t)}\}$. That is we have the sequence

$$X^{(0)}, X^{(1)}, \ldots, X^{(t)}, \ldots$$

such that the probability distribution of $X^{(t)}$ given all the past variables only depends on the very last one $X^{(t-1)}$. This conditional probability is called the transition kernel or Markov kernel $K$, i.e.,

$$X^{(t+1)} | X^{(0)}, X^{(1)}, \ldots, X^{(t)} \sim K(X^{(t)}, X^{(t+1)}).$$

- For a given Markov kernel $K$, there may exist a distribution $f$ such that
$$\int_X K(x, y) f(x)\, dx = f(y).$$
- If $f$ satisfies this equation, we call $f$ a stationary distribution of $K$. What this means is that if $X^{(t)} \sim f$, then $X^{(t+1)} \sim f$ as well.

The theory of Markov chains provides various results about the existence and uniqueness of stationary distributions, but such results are beyond the scope of this course. However, one specific result is that under fairly general conditions that are typically satisfied in practice, if a stationary distribution $f$ exists, then $f$ is the limiting distribution of $\{X^{(t)}\}$ is $f$ for almost any initial value or distribution of $X^{(0)}$. This property is called ergodicity. From a simulation point of view, it means that if a given kernel $K$ produces an ergodic Markov chain with stationary distribution $f$, generating a chain from this kernel will eventually produce simulations that are approximately from $f$.

In particular, a very important result can be derived. For integrable functions $h$, the standard average

$$\frac{1}{M} \sum_{i=1}^{M} h(X^{(t)}) \longrightarrow E_f[h(X)].$$

This means that the LLN lies at the basis of Monte Carlo methods which can be applied in MCMC settings. The result shown above is called the Ergodic Theorem.

Of course, even in applied settings, it should always be confirmed that the Markov chain in question behaves as desired before blindly using MCMC

to perform Bayesian calculations. Again, such theoretical verifications are beyond the scope of this course. Practically speaking, however, the MCMC methods we will discuss do indeed behave nicely in an extremely wide variety on problems.

Now we turn to Gibbs. The name Gibbs sampling comes from a paper by Geman and Geman (1984), which first applied a Gibbs sampler on a Gibbs random field. The name stuck from there. It's actually a special case of something from Markov chain Monte Carlo (MCMC), and more specifically a method called Metropolis-Hastings, which we will hopefully get to. We'll start by studying the simple case of the two-stage sampler and then look at the multi-stage sampler.

### ○ The Two-Stage Gibbs Sampler

The two-stage Gibbs sampler creates a Markov chain from a joint distribution. Suppose we have two random variables $X$ and $Y$ with joint density $f(x, y)$. They also have respective conditional densities $f_{Y|X}$ and $f_{X|Y}$. The two-stage sampler generates a Markov chain $\{(X_t, Y_t)\}$ according to the following steps:

**Algorithm 5.1:** Two-stage Gibbs Sampler
Take $X_0 = x_0$. Then for $t = 1, 2, \ldots,$ generate

1. $X_t \sim f_{X|Y}(\cdot|y_{t-1})$
2. $Y_t \sim f_{Y|X}(\cdot|x_t)$.

As long as we can write down both conditionals (and simulate from them), it is easy to implement the algorithm above.

**Example 5.4:** Bivariate Normal
Consider the bivariate normal model

$$(X, Y) \sim N_2\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Recall the following fact from Casella and Berger (2009): If

$$(X, Y) \sim N_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right),$$

then

$$Y|X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1-\rho^2)\right).$$

Suppose we calculate the Gibbs sampler just given the starting point $(x_0, y_0)$. Since this is a toy example, let's suppose we only care about $X$. Note that we don't really need both components of the starting point, since if we pick $x_0$, we can generate $Y_0$ from $f_{Y|X}(\cdot|x_0)$.

We know that $Y_0 \sim N(\rho x_0, 1-\rho^2)$ and $X_1|Y_0 = y_0 \sim N(\rho y_0, 1-\rho^2)$. Then

$$E[X_1] = E[E[X_1|Y_0]] = \rho x_0$$

and

$$\text{Var}[X_1] = E\text{Var}[X_1|Y_0]] + \text{Var}E[X_1|Y_0]] = 1 - \rho^4.$$

Then

$$X_1 \sim N(\rho^2 x_0, 1-\rho^4).$$

We want the unconditional distribution of $X_2$ eventually. So, we need to update $(X_2, Y_2)$. So we need $Y_1$ so we can generate $Y_1|X_1 = x_1$. Since we only care about $X$, we can use the conditional distribution formula to find that $Y_1|X_1 = x_1 \sim N(\rho x_1, 1-\rho)$. Then using iterated expectation and iterated variance, we can show that

$$X_2 \sim N(\rho^4 x_o, 1-\rho^8).$$

If we keep iterating, we find that

$$X_n \sim N(\rho^{2n} x_o, 1-\rho^{4n}).$$

(To see this, iterate a few times and find the pattern.) What happens as $n \to \infty$?

$$X_n \overset{\text{approx.}}{\sim} N(0,1).$$

**Example 5.5:** Binomial-Beta
Suppose $X|\theta \sim \text{Bin}(n,\theta)$ and $\theta \sim \text{Beta}(a,b)$. Then the joint distribution is

$$f(x,\theta) = \binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{x+a-1}(1-\theta)^{n-x+b-1}.$$

The distribution of $X|\theta$ is given above, and $\theta|X \sim \text{Beta}(x+a, n-x+b)$.

We can implement the Gibbs sampler in R as

```
gibbs_beta_bin <- function(nsim, nn, aa, bb)
{
  xx <- rep(NA,nsim)
  tt <- rep(NA,nsim)

  tt[1] <- rbeta(1,aa,bb)
  xx[1] <- rbinom(1,nn,tt[1])

  for(ii in 2:nsim)
  {
    tt[ii] <- rbeta(1,aa+xx[ii-1],bb+nn-xx[ii-1])
    xx[ii] <- rbinom(1,nn,tt[ii])
  }

  return(list(beta_bin=xx,beta=tt))
}
```

Since $X$ has a discrete distribution, we can use a rootogram to check if the Gibbs sampler performed a good approximation. The rootogram plot is implemented in the library `vcd` in `R`. The following are the commands to generate this rootogram:

```
gibbs_sample <- gibbs_beta_bin(5000,15,3,7)

# Density of a beta-binomial distribution with parameters
# nn: sample size of the binomial
# aa: first parameter of the beta
# bb: second parameter of the beta
dbetabi <- function(xx, nn, aa, bb)
{
  return(choose(nn,xx)*exp(lgamma(aa+xx)-lgamma(aa)+lgamma(nn-xx+bb)-
                           lgamma(bb)-lgamma(nn+aa+bb)+lgamma(aa+bb)))
}

#Rootogram for the marginal distribution of X.
library(vcd)
beta_bin_sample <- gibbs_sample$beta_bin
max_observed <- max(beta_bin_sample)
rootogram(table(beta_bin_sample),5000*dbetabi(0:max_observed,15,3,7),
```

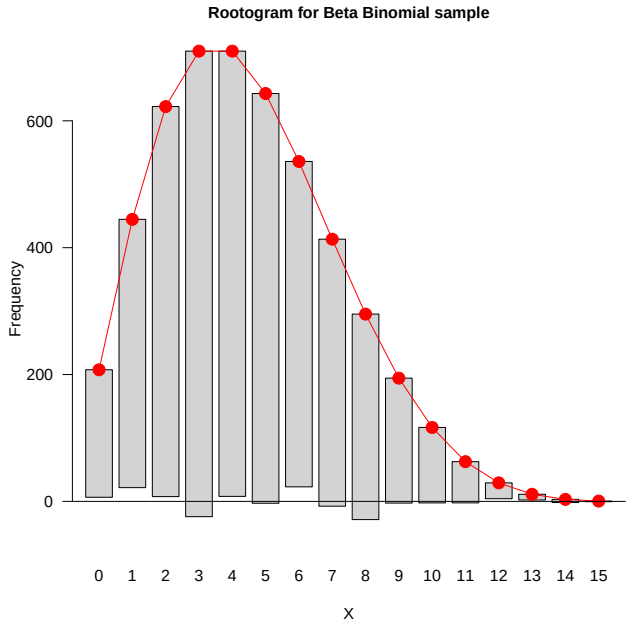



FIGURE 5.5: Rootogram from a Beta-Binomial(15,3,7)

```
  scale="raw",xlab="X",main="Rootogram for Beta Binomial sample")
```

Figure 5.5 presents the rootogram for the Gibbs sample for the Beta-Binomial distribution. Similarly, Figure 5.6 shows the same for the marginal distribution of $\theta$ obtained through the following commands:

```
#Histogram for the marginal distribution of Theta.
beta_sample <- gibbs_sample$beta
hist(beta_sample,probability=TRUE,xlab=expression(theta),
  ylab="Marginal Density", main="Histogram for Beta sample")
curve(dbeta(x,3,7),from=0,to=1,add=TRUE)
```

**Example 5.6:** Consider the posterior on $(\theta, \sigma^2)$ associated with the following model:

$$\begin{aligned} X_i|\theta &\sim N(\theta, \sigma^2),\ i = 1, \ldots, n, \\ \theta &\sim N(\theta_o, \tau^2) \\ \sigma^2 &\sim \text{InverseGamma}(a, b), \end{aligned}$$

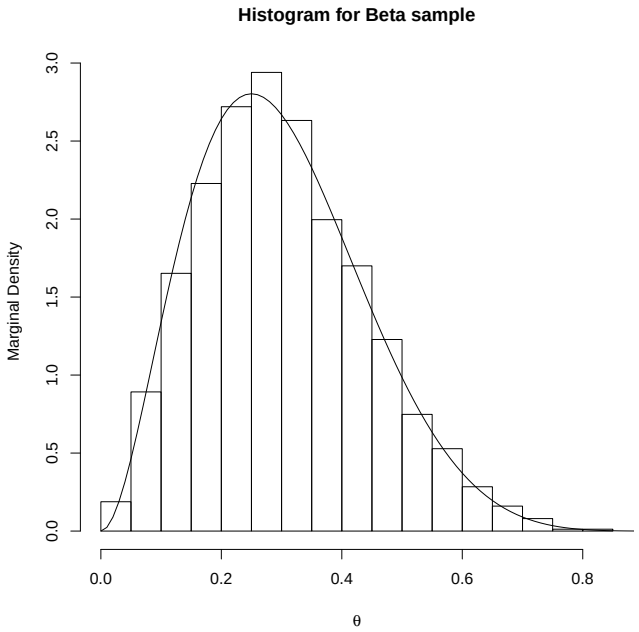



FIGURE 5.6: Histogram for a Beta(3,7)

where $\theta_o, \tau^2, a, b$ known. Recall that $p(\sigma^2) = \dfrac{b^a}{\Gamma(a)} \dfrac{e^{-b/x}}{x^{a+1}}$.

The Gibbs sampler for these conditional distributions can be coded in R as follows:

```
# gibbs_gaussian: Gibbs sampler for marginal of theta|X=xx and sigma2|X=xx
# when Theta ~ Normal(theta0,tau2) and Sigma2 ~ Inv-Gamma(aa,bb) and
# X|Theta=tt,Sigma2=ss ~ Normal(tt,ss)
#
# returns a list gibbs_sample
# gibbs_sample$theta : sample from the marginal distribution of Theta|X=xx
# gibbs_sample$sigma2: sample from the marginal distribution of Sigma2|X=xx

gibbs_gaussian <- function(nsim,xx,theta0,tau2,aa,bb)
{
  nn <- length(xx)
  xbar <- mean(xx)
  RSS <- sum((xx-xbar)^2)
```

```
  post_sigma_shape <- aa + nn/2

  theta <- rep(NA,nsim)
  sigma2 <- rep(NA,nsim)

  sigma2[1] <- 1/rgamma(1,shape=aa,rate=bb)
  ww <- sigma2[1]/(sigma2[1]+nn*tau2)
  theta[1] <- rnorm(1,mean=ww*theta0+(1-ww)*xbar, sd=sqrt(tau2*ww))

  for(ii in 2:nsim)
  {
    new_post_sigma_rate <- (1/2)*(RSS+ nn*(xbar-theta[ii-1])^2) + bb
    sigma2[ii] <- 1/rgamma(1,shape=post_sigma_shape,
      rate=new_post_sigma_rate)

    new_ww <- sigma2[ii]/(sigma2[ii]+nn*tau2)
    theta[ii] <- rnorm(1,mean=new_ww*theta0+(1-new_ww)*xbar,
      sd=sqrt(tau2*new_ww))
  }

  return(list(theta=theta,sigma2=sigma2))
}
```

The histograms in Figure 5.7 for the posterior for $\theta$ and $\sigma^2$ are obtained as follows:

```
library(mcsm)
data(Energy)
gibbs_sample <- gibbs_gaussian(5000,log(Energy[,1]),5,10,3,3)

par(mfrow=c(1,2))
hist(gibbs_sample$theta,xlab=expression(theta~"|X=x"),main="")
hist(sqrt(gibbs_sample$sigma2),xlab=expression(sigma~"|X=x"),main="")
```

### ○ The Multistage Gibbs Sampler

There is a natural extension from the two-stage Gibbs sampler to the general multistage Gibbs sampler. Suppose that for $p > 1$, we can write the

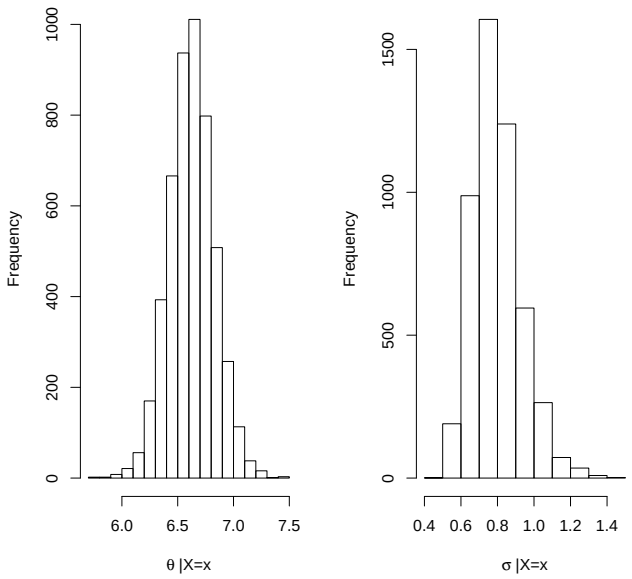



FIGURE 5.7: Histograms for posterior mean and standard deviation.

random variable $\mathbf{X} = (X_1, \ldots, X_p)$, where the $X_i$'s are either unidimensional or multidimensional components. Suppose that we can simulate from corresponding conditional densities $f_1, \ldots, f_p$. That is, we can simulate

$$X_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim f(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

for $i = 1, \ldots, p$. The associated Gibbs sampling algorithm is given by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

**Algorithm 5.2:** The Multistage Gibbs sampler
At iteration $t = 1, 2, \ldots$ given $x^{(t-1)} = (x_1^{(t-1)}, \ldots, x_p^{(t-1)})$, generate

1. $X_1^{(t)} \sim f(x_1 | x_2^{(t-1)}, \ldots, x_p^{(t-1)})$,
2. $X_2^{(t)} \sim f(x_2 | x_1^{(t)}, x_3^{(t-1)} \ldots, x_p^{(t-1)})$,

$\vdots$

p–1. $X_{p-1}^{(t)} \sim f(x_{p-1} | x_1^{(t)}, \ldots, x_{p-2}^{(t)}, x_p^{(t-1)})$,

p. $X_p^{(t)} \sim f(x_p | x_1^{(t)}, \ldots, x_{p-1}^{(t)})$.

The densities $f_1, \ldots, f_p$ are called the full conditionals, and a particular feature of the Gibbs sampler is that these are the only densities used for simulation. Hence, even for high-dimensional problems, all of the simulations may be univariate, which is a major advantage.

**Example 5.7:** (Casella and Robert, p. 207) Consider the following model:

$$\begin{aligned} X_{ij}|\theta_i, \sigma^2 &\overset{ind}{\sim} N(\theta_i, \sigma^2) \qquad 1 \le i \le k, \;\; 1 \le j \le n_i \\ \theta_i|\mu, \tau^2 &\overset{iid}{\sim} N(\mu, \tau^2) \\ \mu|\sigma_\mu^2 &\sim N(\mu_0, \sigma_\mu^2) \\ \sigma^2 &\sim IG(a_1, b_1) \\ \tau^2 &\sim IG(a_2, b_2) \\ \sigma_\mu^2 &\sim IG(a_3, b_3) \end{aligned}$$

The conditional independencies in this example can be visualized by the Bayesian Network in Figure 5.8. Using these conditional independencies, we can compute the complete conditional distributions for each of the variables as

$$\begin{aligned} \theta_i &\sim N\left(\frac{\sigma^2}{\sigma^2 + n_i\tau^2}\mu + \frac{n_i\tau^2}{\sigma^2 + n_i\tau^2}\bar{X}_i, \frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2}\right), \\ \mu &\sim N\left(\frac{\tau^2}{\tau^2 + k\sigma_\mu^2}\mu_0 + \frac{k\sigma_\mu^2}{\tau^2 + k\sigma_\mu^2}\bar{\theta}, \frac{\sigma_\mu^2\tau^2}{\tau^2 + k\sigma_\mu^2}\right), \\ \sigma^2 &\sim IG\left(\sum_i n_i/2 + a_1, (1/2)\sum_{i,j}(X_{i,j} - \theta_i)^2 + b_1\right), \\ \tau^2 &\sim IG\left(k/2 + a_2, (1/2)\sum_i (\theta_i - \mu)^2 + b_2\right), \\ \sigma_\mu^2 &\sim IG\left(1/2 + a_3, 1/2(\mu - \mu_0)^2 + b_3\right), \end{aligned}$$

where $\bar{\theta} = \sum_i n_i\theta_i / \sum_i n_i$.

Running the chain with $\mu_0 = 5$ and $a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = 3$ and chain size 5000, we get the histograms in Figure 5.9.

### ○ Application of the GS to latent variable models

We give an example of Gibbs sampling to an data augmentation example. We look at the example from a genetic linkage analysis. This example is

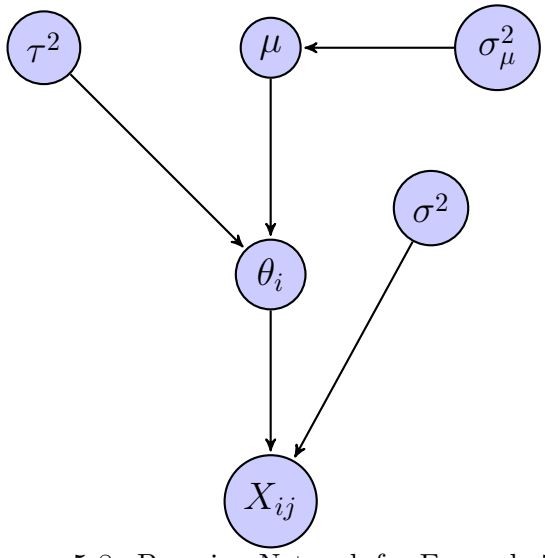



FIGURE 5.8: Bayesian Network for Example 5.7.

given in Rao (1973, pp. 3689) where it is analyzed in a frequentist setting; it was re-analyzed in Dempster, Laird and Rubin (1977), and re-analyzed in a Bayesian framework in Tanner and Wong (1987).

**Example 5.8:** A genetic model specifies that 197 animals are distributed multinomially into four categories, with cell probabilities given by

$$\pi = (1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4).(*)$$

The actual observations are $y = (125, 18, 20, 34)$. We want to estimate $\theta$.

**Biological basis for model**:

Suppose we have two factors, call them $\alpha$ and $\beta$ (say eye color and leg length).

- Each comes at two levels: $\alpha$ comes in levels $A$ and $a$, and $\beta$ comes in levels $B$ and $b$.
- Suppose $A$ is dominant, $a$ is recessive; also $B$ dominant, $b$ recessive.
- Suppose further that $P(A) = 1/2 = P(a)$ [and similarly for the other factor].

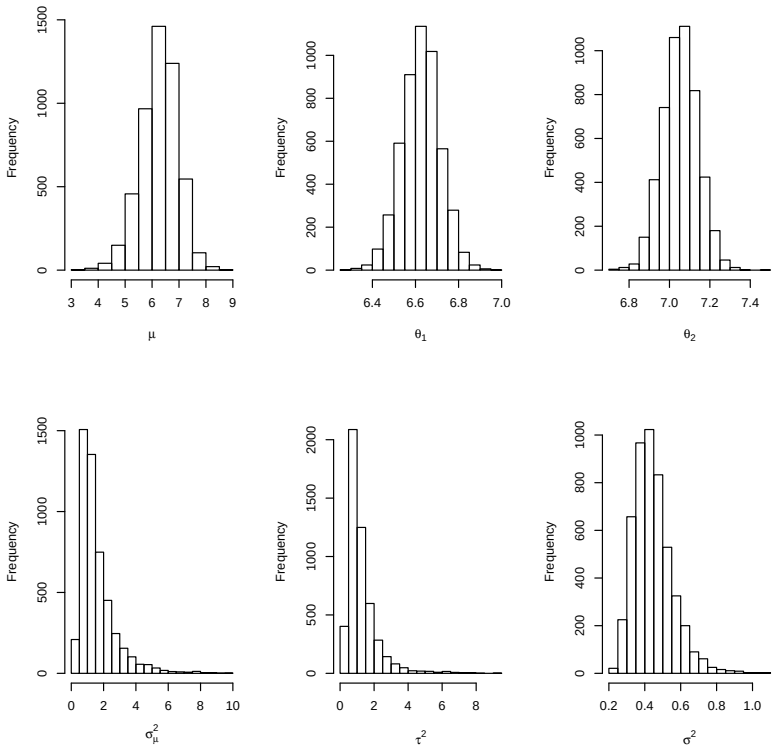



FIGURE 5.9: Histograms for posterior quantities.

- Now suppose that the two factors are related: $P(B|A) = 1-\eta$ and $P(b|A) = \eta$.
- Similarly, $P(B|a) = \eta$ and $P(b|a) = 1-\eta$.

To calculate probability of the phenotypes $AB, Ab, aB$ and $ab$ in an offspring (phenotype is what we actually see in the offspring), we suppose that mother and father are chosen independently from the population, and make following table, involving the genotypes (genotype is what is actually in the genes, and this is not seen).

Then

$$P(\text{Father is AB}) = P(B|A)P(B) = \frac{1}{2}(1-\eta).$$

$$P(\text{Mother is AB}) = P(B|A)P(B) = \frac{1}{2}(1-\eta).$$

$$P(\text{O.S. is AB}) = P(B|A)P(B) = \frac{1}{4}(1-\eta)^2.$$

Note: $\eta = 1/2$ means no linkage and people like to estimate $\eta$.

TABLE 5.1: default

| | AB | Ab | aB | ab |
|---|---|---|---|---|
| AB | $\frac{1}{4}(1-\eta)^2$ | $\frac{1}{4}(1-\eta)\eta$ | $\frac{1}{4}(1-\eta)\eta$ | $\frac{1}{4}(1-\eta)^2$ |
| Ab | $\frac{1}{4}(1-\eta)\eta$ | | | |
| aB | | | | |
| ab | | | | |

There are 9 cases where we would see the phenotype AB and adding up their probabilities, we get $\frac{3-2\eta+\eta^2}{4}$. You can find similar probabilities for the other phenotypes. Writing

$$\frac{(3-2\eta+\eta^2)}{4} = \frac{1}{2} + \frac{1-2\eta+\eta^2}{4}$$

and letting $\theta = (1-\eta)^2$, we find the model specified in (*).

**What now?**
Suppose we put the prior Beta$(a, b)$ on $\theta$. How do we get the posterior?

Here is one method, using the Gibbs sampler.

Split first cell into two cells, one with probability 1/2, the other with probability $\theta/4$.

Augment the data into a 5-category multinomial, call it $X$, where $X_1$ is Bernoulli with parameter 1/2. Now consider $p_{data}(X_1|\theta)$. Will run a Gibbs sampler of length 2:

- The conditional distribution of $X_1 \mid \theta$ (and given the data) is $Bin(125, \frac{1/2}{1/2+\theta/4})$.
- Given the data, conditional on X1 the model is simply a binomial with $n = 197 - X_1$, and probability of success $\theta$, and data consisting of $(125 - X_1 + X_5)$successes, $(X_3 + X_4)$ failures
- Thus, conditional distribution of $\theta \mid X_1$ and the data is

$$\text{Beta}(a + 125 - X_1 - X_5, b + X_3 + X_4).$$

R Code to implement G.S.:

```
set.seed(1)
a <- 1; b <- 1
z <- c(125,18,20,34)
x <- c(z[1]/2, z[1]/2, z[2:4])
nsim <- 50000 # runs in about 2 seconds on 3.8GHz P4
theta <- rep(a/(a+b), nsim)
for (j in 1:nsim)
{
  theta[j] <- rbeta(n=1, shape1=a+125-x[1]+x[5],
                    shape2=b+x[3]+x[4])
  x[1] <- rbinom(n=1, z[1], (2/(2+theta[j])))
}
mean(theta) # gives 0.623
pdf(file="post-dist-theta.pdf",
               horiz=F, height=5.0, width=5.0)
plot(density(theta), xlab=expression(theta), ylab="",
     main=expression(paste("Post Dist of ", theta)))
dev.off()
eta <- 1 - sqrt(theta) # Variable of actual interest
plot(density(eta))
```

```
sum(eta > .4)/nsim # gives 0
```

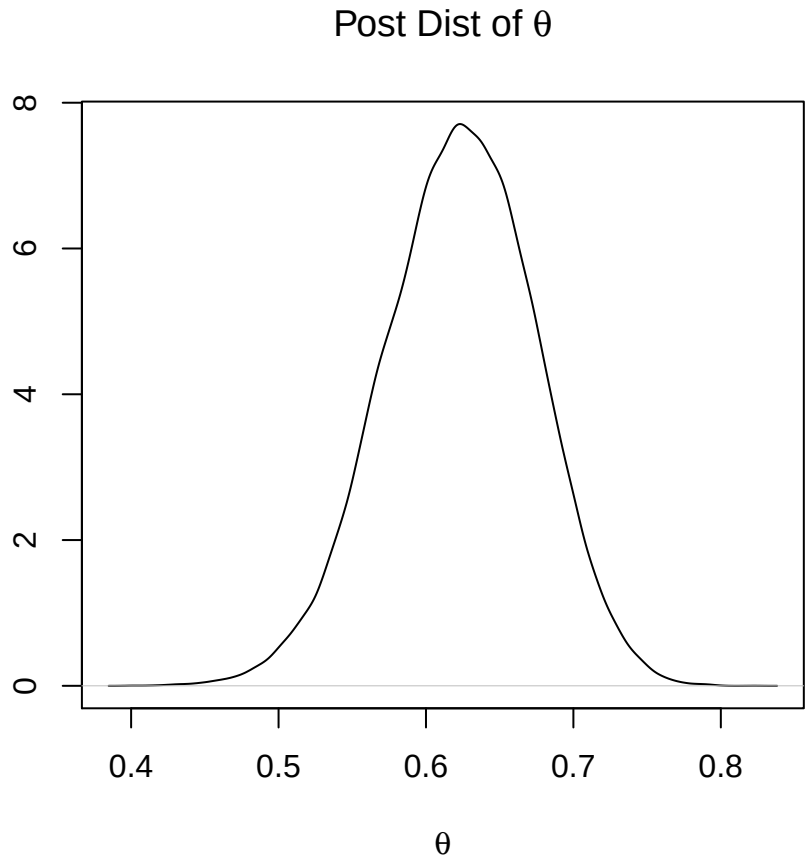



FIGURE 5.10: **Posterior Distribution of $\theta$ for Genetic Linkage**

## 5.3 MCMC Diagnostics

We will want to check any chain that we run to assess any lack of convergence.

The adequate length of a run will depend on

- a burn-in period (debatable topic).
- mixing rate.
- variance of quantity we are monitoring.

Quick checks:

- trace plots: a times series plot of the parameters of interest; indicates how quickly the chain is mixing of failure to mix.
- Autocorrelations plots.
- Plots of log posterior densities – used mostly in high dimensional problems.
- Multiple starting points – diagnostic to attempt to handle problems when we obtain different estimates when we start with multiple (different) starting values.

**Definition**: An autocorrelation plot graphically measures the correlation between $X_i$ and each $X_{k+i}$ variable in the chain.

- The Lag-k correlation is the $\text{Corr}(X_i, X_{k+i})$.
- By looking at autocorrelation plots of parameters that we are interested in, we can decide how much to thin or subsample our chain by.
- Then rerun Gibbs sampler using new thin value.

For a real data example that I'm working on:

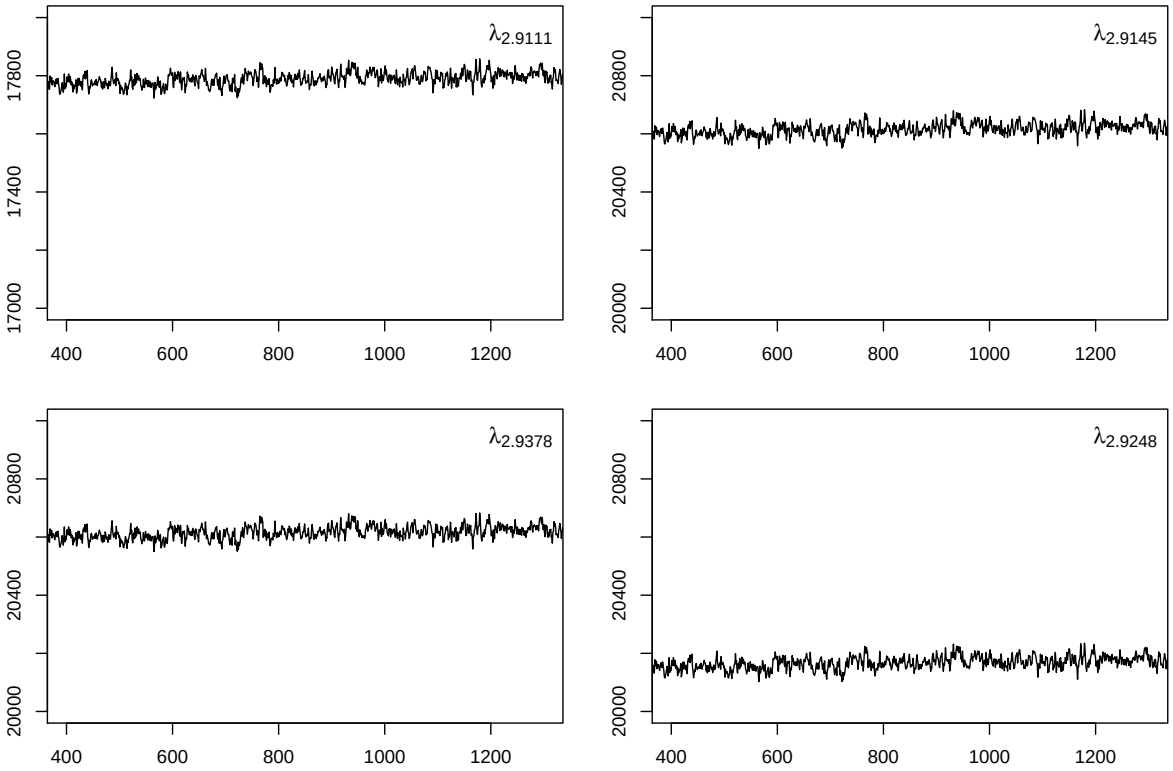



FIGURE 5.11: Trace Plot for RL Example

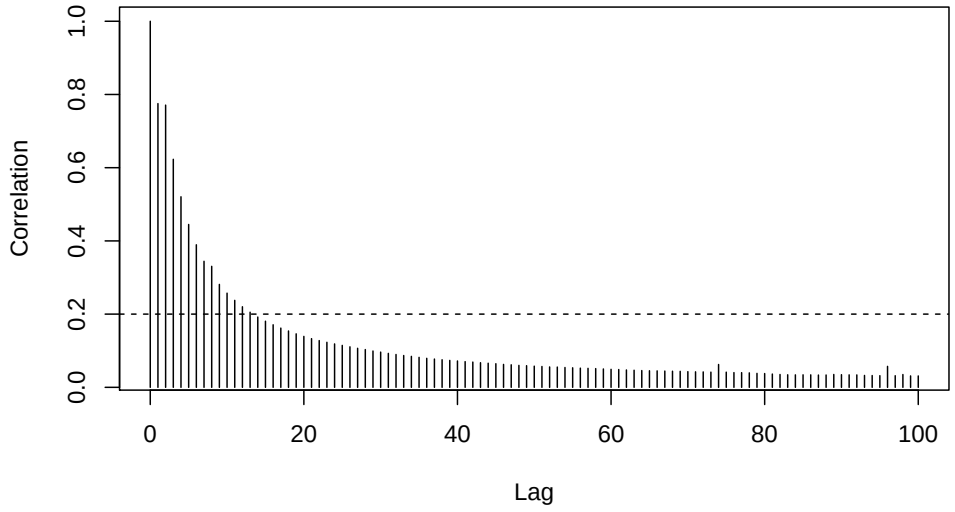



FIGURE 5.12: Max Autocorrelation Plot for RL Example

**Multiple Starting Points**: Can help determine if burn-in is long enough.

- Basic idea: want to estimate the mean of a parameter $\theta$.
- Run chain 1 starting at $x_o$. Estimate the mean to be $10 \pm 0.1$.
- Run chain 2 starting at $x_1$. Estimate the mean to be $11 \pm 0.1$.
- Then we know that the effort of the starting point hasn't been forgotten.
- Maybe the chain hasn't reached the area of high probability yet and need to be run for longer?
- Try running multiple chains.

**Gelman-Rubin**

- Idea is that if we run several chains, the behavior of the chains should be basically the same.

- Check informally using trace plots.
- Check using the Gelman-Rubin diagnostic – but can fail like any test.
- Suggestions – Geweke – more robust when normality fails.

## 5.4 Theory and Application Based Example

### ○ PlA2 Example

Twelve studies run to investigate potential link between presence of a certain genetic trait and risk of heart attack. Each was case-control, and considered a group of individuals with coronary heart disease and another group with no history of heart disease. For each study i (i = 1, . . . , 12) the proportion having the genetic trait in each group was noted and a log odds ratio $\hat{\psi}_i$ was calculated, together with a standard error $\sigma_i$. Results are summarized in table below (data from Burr et al. 2003).

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\hat{\psi}_i$ | 1.06 | -0.10 | 0.62 | 0.02 | 1.07 | -0.02 |
| $\sigma_i$ | 0.37 | 0.11 | 0.22 | 0.11 | 0.12 | 0.12 |

| $i$ | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|
| $\hat{\psi}_i$ | -0.12 | -0.38 | 0.51 | 0.00 | 0.38 | 0.40 |
| $\sigma_i$ | 0.22 | 0.23 | 0.18 | 0.32 | 0.20 | 0.25 |

Setup:

- Twelve studies were run to investigate the potential link between presence of a certain genetic trait and risk of heart attack.
- Each study was case-control and considered a group of individuals with coronary heart disease and another group with no history of coronary heart disease.
- For each study $i$ ($i = 1, \cdots, 12$) the proportion having the genetic trait in each group was recorded.
- For each study, a log odds ratio, $\hat{\psi}_i$, and standard error, $\sigma_i$, were calculated.

Let $\psi_i$ represent the true log odds ratio for study $i$. Then a typical hierarchical model would look like:

$$\begin{aligned}\hat{\psi}_i \mid \psi_i &\overset{ind}{\sim} N(\psi_i, \sigma_i^2)\ i = 1, \ldots, 12\\ \psi_i \mid \mu, \tau &\overset{iid}{\sim} N(\mu, \tau^2)\ i = 1, \ldots, 12\\ (\mu, \tau) &\sim \nu.\end{aligned}$$

From this, the likelihood is

$$L(\mu, \tau) = \int \ldots \int \prod_{i=1}^{12} N_{\psi_i, \sigma_i}(\hat{\psi}_i) \prod_{i=1}^{12} N_{\mu,\tau}(\psi_i)\ d\psi_1 \ldots d\psi_{12}.$$

The posterior can be written as (as long as $\mu$ and $\tau$ have densities), as

$$\pi(\mu, \tau \mid \hat{\psi}_i) = c^{-1}\left[\int \ldots \int \prod_{i=1}^{12} N_{\psi_i, \sigma_i}(\hat{\psi}_i) \prod_{i=1}^{12} N_{\mu,\tau}(\psi_i)\ d\psi_1 \ldots d\psi_{12}\right] p(\mu, \tau).$$

Supppose we take $\nu$ = "Normal/Inverse Gamme prior." Then conditional on $\tau, \mu \sim N(c, d\tau^2)$, $\gamma = 1/\tau^2 \sim Gamma(a, b)$.

Remark: The reason for taking this prior is that it is conjugate for the normal distribution with both mean and variance unknown (that is, it is conjugate for the model in which the $\psi_i$'s are observed).

We will use the notation NIG(a, b, c, d) to denote this prior. Taking a = .1, b = .1, c = 0, and d = 1000 gives a flat prior.

- If we are frequentists, then we need to calculated the likelihood

$$L(\mu, \tau) = \int \ldots \int \prod_{i=1}^{12} N_{\psi_i, \sigma_i}(\hat{\psi}_i) \prod_{i=1}^{12} N_{\mu,\tau}(\psi_i)\ d\psi_1 \ldots d\psi_{12}.$$

- If we are Bayesians, we need to calculate the likelihood and in addition we need to calculate the normalizing constant in order to find the posterior

$$pi(\mu, \tau \mid \hat{\psi}_i) = \frac{L(\mu, \tau)p(\mu, \tau)}{\int L(\mu, \tau)p(\mu, \tau)d\mu d\tau}$$

- Neither above is easy to do.

We have a choice:

- Select a model that doesn't fit the data well but gives answers that are easy to obtain, i.e. in closed form.
- Select a model that is appropriate for the data but i computationally difficult to deal with.

MCMC methods often allow us (in many cases) to make the second choice.

Going back to the example and fitting a model:

**Recall the general model**:

$$\begin{aligned}\hat{\psi}_i \mid \psi_i &\overset{ind}{\sim} N(\psi_i, \sigma_i^2)\ i = 1, \ldots, 12\\ \psi_i \mid \mu, \tau &\overset{iid}{\sim} N(\mu, \tau^2)\ i = 1, \ldots, 12\\ (\mu, \tau) &\sim NIG(a, b, c, d).\end{aligned}$$

Then the posterior of $(\mu, \tau)$ is NIG(a',b',c',d'), with

$$a' = a + n/2 \qquad b' = b + \frac{1}{2}\sum_i (X_i - \bar{X})^2 + \frac{n(\bar{X} - c)^2}{2(1 + nd)}$$

and

$$c' = \frac{c + nd\bar{X}}{nd + 1} \qquad d' = \frac{1}{n + d^{-1}}.$$

This means that

$$\mu \mid \tau^2, y \sim N(c', d')$$

and

$$\tau^2 \mid y \sim \text{InverseGamma}(a', b')$$

**Implementing the Gibbs sampler**:

Want the posterior distribution of $(\mu, \tau)$.

- In order to clarify what we are doing, we use the notation that subscripting a distribution by a random variable denotes conditioning.
- Thus, if $U$ and $V$ are two random variables, $L(U|V)$ and $L_V(U)$ will both denote the conditional distribution of $U$ given $V$.

We want to find $L_{\hat{\psi}}(\mu, \tau, \psi)$. We'll run a Gibbs sampler of length 2:

- Given $(\mu, \tau)$, the $\psi$'s are independent. The conditional distribution of $\psi$ given $\hat{\psi}$ is the conditional distribution of $\psi$ given only $\hat{\psi}$. This conditional distribution is given by a standard result for the conjugate normal/normal situation: it is $N(\mu', \tau'^2)$, where

$$\mu' = \frac{\sigma_i^2 \mu + \tau^2 \hat{\psi}_i}{\sigma_i^2 + \tau^2} \qquad \tau'^2 = \frac{\sigma_i^2 \tau^2}{\sigma_i^2 + \tau^2}$$

- Given the ($\psi$'s, the data) are superfluous, i.e. $L_{\hat{\psi}}(\mu, \tau \mid \psi) = L(\mu, \tau \mid \psi)$. This conditional distribution is given by the conjugacy of the Normal / Inverse gamma prior: $L(\mu, \tau \mid \psi) = \text{NIG}(a', b', c', d')$, where

$$a' = a + n/2 \qquad b' = b + \frac{1}{2}\sum_i (\psi_i - \bar{\psi})^2 + \frac{n(\bar{\psi} - c)^2}{2(1 + nd)}$$

and

$$c' = \frac{c + nd\bar{\psi}}{nd + 1} \qquad d' = \frac{1}{n + d^{-1}}.$$

This gives us a sequence $\mu, \tau, \psi_1, \ldots, \psi_n; g = 1, ..., G$, from $L_{\hat{\psi}}(\mu, \tau, \psi)$. If were interested in, e.g., the posterior distribution of $\mu$, we just retain the first coordinate in the sequence.

**Specific Example for PlA2 data**
Our proposed hierarchical model is

$$\hat{\psi}_i \mid \psi_i \overset{ind}{\sim} N(\psi_i, \sigma_i^2) \quad i = 1, \cdots, 12$$

$$\psi_i \mid \mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2) \quad i = 1, \cdots, 12$$

$$\mu \mid \tau^2 \sim N(0, 1000\tau^2)$$

$$\gamma = 1/\tau^2 \sim \text{Gamma}(0.1, 0.1)$$

Why is a normal prior taken? It's conjugate for the normal distribution with the mean and variance known. The two priors above with the chosen hyperparameters result in noninformative hyperpriors.

```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file model.txt contains

\begin{verbatim}
model {
for (i in 1:N) {
psihat[i] ~ dnorm(psi[i],1/(sigma[i])^2)
psi[i] ~ dnorm(mu,1/tau^2)
}
mu ~ dnorm(0,1/(1000*tau^2))
tau <- 1/sqrt(gam)
gam ~ dgamma(0.1,0.1)
}
```

Note: In BUGS, use dnorm(mean,precision), where precision = 1/variance.

```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file data.txt contains

\begin{verbatim}
"N" <- 12
"psihat" <-  c(1.055, -0.097, 0.626, 0.017, 1.068,
-0.025, -0.117, -0.381, 0.507, 0, 0.385, 0.405)
"sigma" <- c(0.373, 0.116, 0.229, 0.117, 0.471,
0.120, 0.220, 0.239, 0.186, 0.328, 0.206, 0.254)
```

The file inits_1.txt contains

```
".RNG.name" <- "base::Super-Duper"
".RNG.seed" <- 12
"psi" <- c(0,0,0,0,0,0,0,0,0,0,0,0)
"mu" <- 0
"gam" <- 1

%\frame[containsverbatim]{
```

```
%\frametitle{PlA2 Example}

The file script.txt contains
\small
\begin{verbatim}
model clear
data clear
model in "model"
data in "data"
compile, nchains(2)
inits in "inits1", chain(1)
inits in "inits2", chain(2)
initialize
update 10000
monitor mu
monitor psi
monitor gam
update 100000
coda *, stem(CODA1)
coda *, stem(CODA2)
```

Now, we read in the coda files into R from the current directory and continue our analysis. The first part of our analysis will consist of some diagnostic procedures.

We will consider

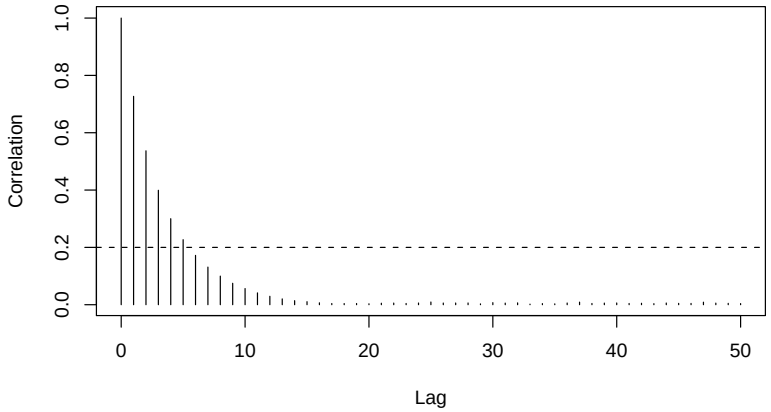
- Autocorrelation Plots
- Trace Plots
- Gelman-Rubin Diagnostic
- Geweke Diagnostic

**Definition**: An autocorrelation plot graphically measures the correlation between $X_i$ and each $X_{k+i}$ variable in the chain.

- The Lag-k correlation is the $\text{Corr}(X_i, X_{k+i})$.

- By looking at autocorrelation plots of parameters that we are interested in, we can decide how much to thin or subsample our chain by.
- We can rerun our JAGS script using our thin value.

We take the thin value to be the first lag whose correlation $\leq 0.2$. For this plot, we take a thin of 2. We will go back and rerun our JAGS script and skip every other value in each chain. After thinning, we will proceed with other diagnostic procedures of interest.

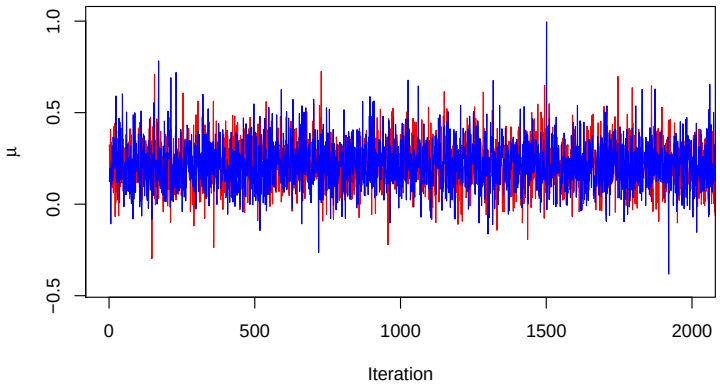



```
%\frame[containsverbatim]{
%\frametitle{PlA2 Example}

The file script\_thin.txt contains
\small
\begin{verbatim}
model clear
data clear
model in "model"
data in "data"
compile, nchains(2)
inits in "inits1", chain(1)
inits in "inits2", chain(2)
initialize
update 10000
```

```
monitor mu, thin(6)
monitor psi, thin(6)
monitor gam, thin(6)
update 100000
coda *, stem(CODA1_thin)
coda *, stem(CODA2_thin)
```

**Definition**: A trace plot is a time series plot of the parameter, say $\mu$, that we monitor as the Markov chain(s) proceed(s).




**Definition**: The Gelman-Rubin diagnostic tests that burn-in is adequate and requires that multiple starting points be used.

To compute the G-R statistic, we must

- Run two chains in JAGS using two different sets of initial values (and two different seeds).
- Load coda package in R and run gelman.diag(mcmc.list(chain1,chain2)).

How do we interpret the Gelman-Rubin Diagnostic?

- If the chain has reached convergence, the G-R test statistic R $\approx$ 1. We conclude that burn-in is adequate.

- Values above 1.05 indicate lack of convergence.

**Warning**: The distribution of R under the null hypothesis is essentially an F distribution. Recall that the F-test for comparing two variances is not robust to violations of normality. Thus, we want to be cautious in using the G-R diagnostic.

```
%\frame[containsverbatim]{
%\frametitle{Gelman-Rubin Diagnostic}

Doing this in R, for the PlA-2 example, we find

\begin{verbatim}
        Point est. 97.5% quantile
mu               1              1
psi[1]           1              1
psi[2]           1              1
...
psi[11]          1              1
psi[12]          1              1
gam              1              1
```

Since 1 is in all the 95% CI, we can conclude that we have not failed to converge.

Suppose $\mu$ is the parameter of interest.

**Main Idea**: If burn-in is adequate, the mean of the posterior distribution of $\mu$ from the first half of the chain should equal the mean from the second half of the chain.

To compute the Geweke statistic, we must

- Run a chain in JAGS along with a set of initial values.
- Load the coda package in R and run geweke.diag(mcmc.list(chain)).

- The Geweke statistic asymptotically has a standard normal distribution, so if the values from R are outside -2.5 or 2.5, this indicates nonstationarity of chain and that burn-in is not sufficient.
- Using the Geweke diagnostic on the PlA2 data indicates that burn-in of 10,000 is sufficient (the largest absolute Z-score is 1.75).
- Observe that the Geweke diagnostic does not require multiple starting points as Gelman-Rubin does.
- The Geweke statistic (based on a T-test) is robust against violations of normality so the Geweke test is preferred to Gelman-Rubin.

Using Gelman-Rubin and Geweke we have shown that burn-in is "sufficient."

- We can look at summary statistics such as means, standard errors, and credible intervals using the summary function.
- We can use kernel density functions in R to estimate posterior distributions that we are interested in using the density function.

| | Post Mean | Post SD | Post Naive SE |
|---|---|---|---|
| $\mu$ | 0.217272 | 0.127 | 0.0009834 |
| $\psi_1$ | 0.594141 | 0.2883 | 0.0022334 |
| $\psi_2$ | -0.062498 | 0.1108 | 0.0008583 |
| $\psi_3$ | 0.490872 | 0.2012 | 0.0015588 |
| $\psi_4$ | 0.040284 | 0.1118 | 0.0008658 |
| $\psi_5$ | 0.51521 | 0.3157 | 0.0024453 |
| $\psi_6$ | 0.003678 | 0.114 | 0.0008831 |
| $\psi_7$ | -0.015558 | 0.1883 | 0.0014586 |
| $\psi_8$ | -0.175852 | 0.2064 | 0.0015988 |
| $\psi_9$ | 0.433525 | 0.1689 | 0.0013084 |
| $\psi_{10}$ | 0.101912 | 0.2423 | 0.0018769 |
| $\psi_{11}$ | 0.332775 | 0.1803 | 0.0013965 |
| $\psi_{12}$ | 0.331466 | 0.2107 | 0.0016318 |
| $\gamma$ | 10.465411 | 6.6611 | 0.051596 |

The posterior of $\mu \mid$ data is

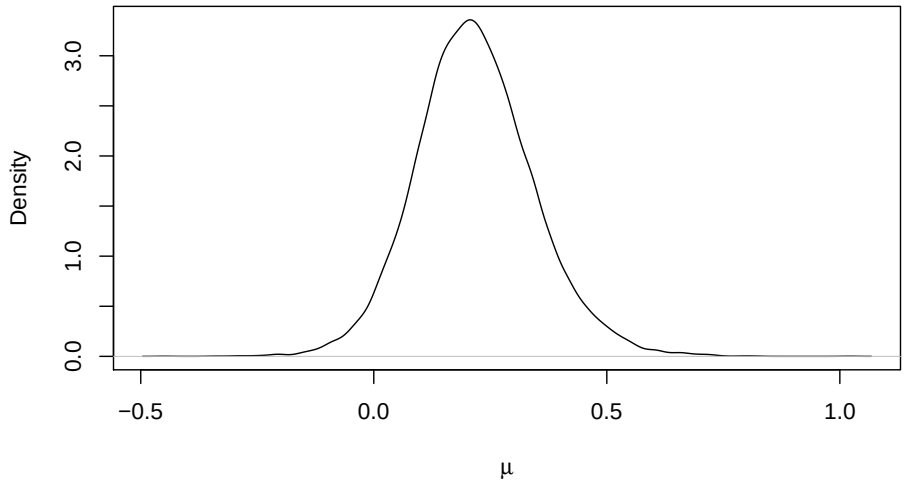
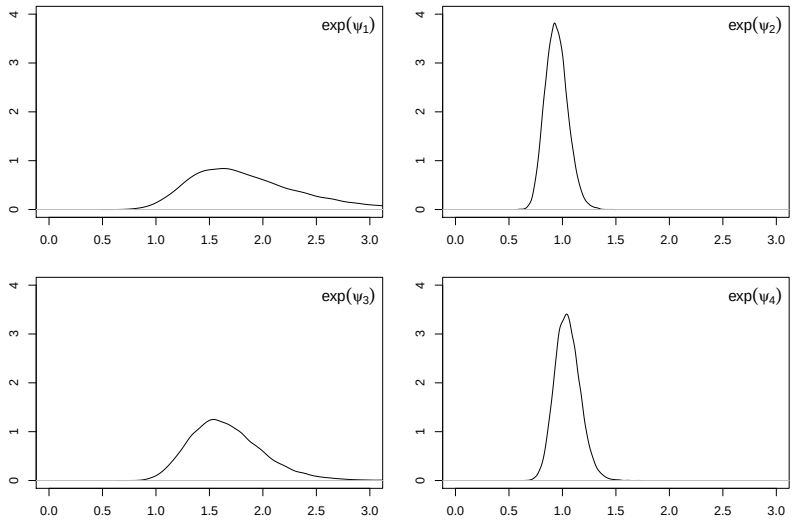



FIGURE 5.13: Posterior of $\mu \mid$ data

Alternatively, we can estimate the conditional distributions of $\exp(\psi_i)$'s given the data. A few are shown below.

- So, here we're looking at the odds ratio's of the prob of getting heart disease given you have the genetic trait over the prob of not getting heart disease given you have the trait. Note that all estimates are pulled toward the mean showing a Bayesian Stein effect.
- This is the odds ratio of having a heart attack for those who have the genetic trait versus those who don't (looking at study i).




```
%\frame[containsverbatim]{
Moreover, we could have just have easily done this analysis in WinBUGS. Below is the
\begin{verbatim}
model{
for (i in 1:N) {
psihat[i] ~ dnorm(psi[i],rho[i])
psi[i] ~ dnorm(mu,gam)
rho[i] <- 1/pow(sigma[i],2)
}

mu ~ dnorm(0,gamt)
gam ~ dgamma(0.1,0.1)
gamt <- gam/1000
}
```

Finally, we can either run the analysis using WinBUGS or JAGS and R. I will demonstrate how to do this using JAGS for this example. I have included the basic code to run this on a Windows machine via WinBUGS. Both methods yield essentially the same results.

To run WinBUGS within `R`, you need the following:

- Load the `R2WinBUGS` library.

- Read in data and format it as a `list()`.
- Format intital values as a `list()`.
- Format the unknown parameters using `c()`.
- Run the `bugs()` command to open/run WinBUGS.
- Read in the G.S. values using `read.coda()`.

```
\scriptsize
\begin{verbatim}
setwd("C:/Documents and Settings/Tina Greenly
/Desktop/beka_winbugs/novartis/pla2")
library(R2WinBUGS)
pla2 <- read.table("pla2_data.txt",header=T)
attach(pla2)
names(pla2)
N<-length(psihat)
data <- list("psihat", "sigma", "N")

inits1 <- list(psi = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), mu = 0, gam = 1)
inits2 <- list(psi = c(2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2), mu = 1, gam = 2)
inits = list(inits1, inits2)
parameters <- c("mu", "psi", "gam")
pla2.sim <- bugs(data, inits, parameters,
  "pla2.bug", n.chains=2, n.iter = 110000,
  codaPkg=T,debug=T,n.burnin = 10000,n.thin=1,bugs.seed=c(12,13),
working.directory="C:/Documents and Settings/Tina Greenly/
Desktop/beka_winbugs/novartis/pla2")
detach(pla2)
coda1 = read.coda("coda1.txt","codaIndex.txt")
coda2 = read.coda("coda2.txt","codaIndex.txt")
```

## 5.5 Metropolis and Metropolis-Hastings

The Metropolis-Hastings algorithm is a general term for a family of Markov chain simulation methods that are useful for drawing samples from Bayesian posterior distributions. The Gibbs sampler can be viewed as a special case

of Metropolis-Hastings (as well will soon see). Here, we present the basic Metropolis algorithm and its generalization to the Metropolis-Hastings algorithm, which is often useful in applications (and has many extensions).

Suppose we can sample from $p(\theta|y)$. Then we could generate

$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{iid}{\sim} p(\theta|y)$$

and obtain Monte Carlo approximations of posterior quantities

$$E[g(\theta)|y] \to 1/S \sum_{i=1}^{S} g(\theta^{(i)}).$$

But what if we cannot sample directly from $p(\theta|y)$? The important concept here is that we are able to construct a large collection of $\theta$ values (rather than them being iid, since this most certain for most realistic situations will not hold). Thus, for any two different $\theta$ values $\theta_a$ and $\theta_b$, we need

$$\frac{\#\theta' s \text{ in the collection } = \theta_a}{\#\theta' s \text{ in the collection } = \theta_b} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}.$$

How might we intuitively construct such a collection?

- Suppose we have a working collection $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ and we want to add a new value $\theta^{(s+1)}$.
- Consider adding a value $\theta^*$ which is nearby $\theta^{(s)}$.
- Should we include $\theta^*$ or not?
- If $p(\theta^*|y) > p(\theta^{(s)}|y)$, then we want more $\theta^*$'s in the set than $\theta^{(s)}$'s.
- But if $p(\theta^*|y) < p(\theta^{(s)}|y)$, we shouldn't necessarily include $\theta^*$.

Based on the above, perhaps our decision to include $\theta^*$ or not should be based upon a comparison of $p(\theta^*|y)$ and $p(\theta^{(s)}|y)$. We can do this by computing r:

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y \mid \theta^*)p(\theta^*)}{p(y \mid \theta^{(s)})p(\theta^{(s)})}.$$

Having computed $r$, what should we do next?

- If $r > 1$ (intuition): Since $\theta^{(s)}$ is already in our set, we should include $\theta^*$ as it has a higher probability than $\theta^{(s)}$.

  (procedure): Accept $\theta^*$ into our set and let $\theta^{(s+1)} = \theta^*$.

- If $r < 1$ (intuition): The relative frequency of $\theta$-values in our set equal to $\theta^*$ compared to those equal to $\theta^{(s)}$ should be

  $$\frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = r.$$

  This means that for every instance of $\theta^{(s)}$, we should only have a fraction of an instance of a $\theta^*$ value.

  (procedure): Set $\theta^{(s+1)}$ equal to either $\theta^*$ or $\theta^{(s)}$ with probability $r$ and $1 - r$ respectively.

This is basic intuition behind the Metropolis (1953) algorithm. More formally, it

- It proceeds by sampling a proposal value $\theta^*$ nearby the current value $\theta^{(s)}$ using a *symmetric proposal distribution* $J(\theta^* \mid \theta^{(s)})$.
- What does symmetry mean here? It means that $J(\theta_a \mid \theta_b) = J(\theta_b \mid \theta_a)$. That is, the probability of proposing $\theta^* = \theta_a$ given that $\theta^{(s)} = \theta_b$ is equal to the probability of proposing $\theta^* = \theta_b$ given that $\theta^{(s)} = \theta_a$.
- Symmetric proposals include:

  $$J(\theta^* \mid \theta^{(s)}) = \text{Uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta)$$

  and

  $$J(\theta^* \mid \theta^{(s)}) = \text{Normal}(\theta^{(s)}, \delta^2).$$

The Metropolis algorithm proceeds as follows:

1. Sample $\theta^* \sim J(\theta \mid \theta^{(s)})$.
2. Compute the acceptance ratio (r):

   $$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y \mid \theta^*)p(\theta^*)}{p(y \mid \theta^{(s)})p(\theta^{(s)})}.$$

3. Let

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with prob min(r,1)} \\ \theta^{(s)} & \text{otherwise.} \end{cases}$$

Remark: Step 3 can be accomplished by sampling $u \sim \text{Uniform}(0,1)$ and setting $\theta^{(s+1)} = \theta^*$ if $u < r$ and setting $\theta^{(s+1)} = \theta^{(s)}$ otherwise.

**Example 5.9:** Metropolis for Normal-Normal
Let's test out the Metropolis algorithm for the conjugate Normal-Normal model with a known variance situation.

That is let

$$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \text{Normal}(\theta, \sigma^2)$$
$$\theta \sim \text{Normal}(\mu, \tau^2).$$

Recall that the posterior of $\theta$ is Normal$(\mu_n, \tau_n^2)$, where

$$\mu_n = \bar{x}\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} + \mu\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$$

and

$$\tau_n^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}.$$

Suppose (taken from Hoff, 2009), $\sigma^2 = 1$, $\tau^2 = 10$, $\mu = 5$, $n = 5$, and $y = (9.37, 10.18, 9.16, 11.60, 10.33)$. For these data, $\mu_n = 10.03$ and $\tau_n^2 = 0.20$.

Suppose that for some ridiculous reason we cannot come up with the posterior distribution and instead we need the Metropolis algorithm to approximate it (please note how incredible silly this example is and it's just to illustrate the method).

Based on this model and prior, we need to compute the acceptance ratio $r$

$$r = \frac{p(\theta^*|x)}{p(\theta^{(s)}|x)} = \frac{p(x|\theta^*)p(\theta^*)}{p(x|\theta^{(s)})p(\theta^{(s)})} = \left(\frac{\prod_i \text{dnorm}(x_i, \theta^*, \sigma)}{\prod_i \text{dnorm}(x_i, \theta^{(s)}, \sigma)}\right)\left(\frac{\prod_i \text{dnorm}(\theta^*, \mu, \sigma)}{\prod_i \text{dnorm}(\theta^{(s)}, \mu, \sigma)}\right)$$

In many cases, computing the ratio $r$ directly can be numerically unstable, however, this can be modified by taking $\log r$.

This results in

$$\log r = \sum_i \Big[\log \text{dnorm}(x_i, \theta^*, \sigma) - \log \text{dnorm}(x_i, \theta^{(s)}, \sigma)\Big]$$
$$+ \sum_i \Big[\log \text{dnorm}(\theta^*, \mu, \sigma) - \log \text{dnorm}(\theta^{(s)}, \mu, \sigma)\Big].$$

Then a proposal is accepted if $\log u < \log r$, where $u$ is sample from the Uniform(0,1).

The R-code below generates 10,000 iterations of the Metropolis algorithm stating at $\theta^{(0)} = 0.$ and using a normal proposal distribution, where

$$\theta^{(s+1)} \sim \text{Normal}(\theta^{(s)}, 2).$$

Below is `R`-code for running the above model. Figure 5.14 shows a trace plot for this run as well as a histogram for the Metropolis algorithm compared with a draw from the true normal density. From the trace plot, although the value of $\theta$ does not start near the posterior mean of 10.03, it quickly arrives there after just a few iterations. The second plot shows that the empirical distribution of the simulated values is very close to the true posterior distribution.

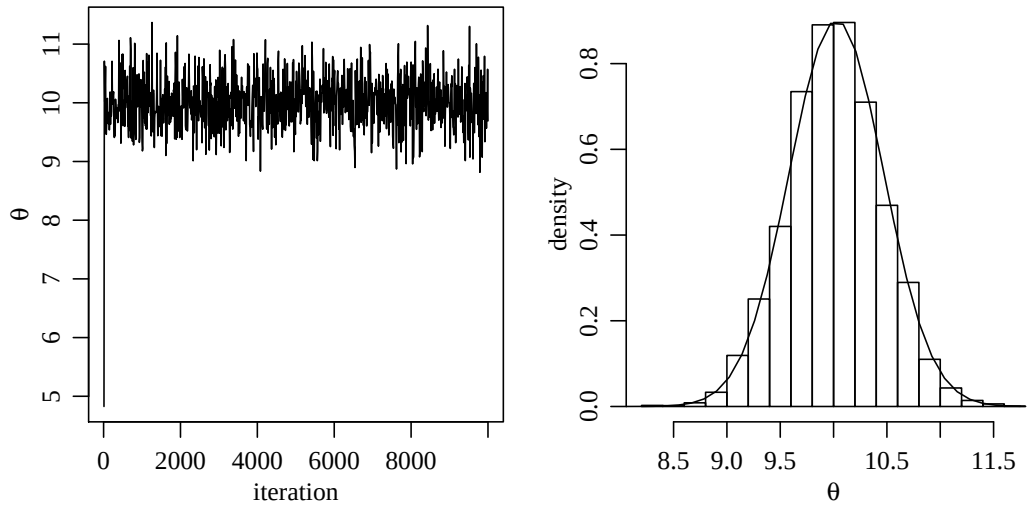



FIGURE 5.14: Results from the Metropolis sampler for the normal model.

```
## initialing values for normal-normal example and setting seed
# MH algorithm for one-sample normal problem with known variance

s2<-1
t2<-10 ; mu<-5; set.seed(1); n<-5; y<-round(rnorm(n,10,1),2)
mu.n<-( mean(y)*n/s2 + mu/t2 )/( n/s2+1/t2)
t2.n<-1/(n/s2+1/t2)

####metropolis part####
y<-c(9.37, 10.18, 9.16, 11.60, 10.33)
##S = total num of simulations
theta<-0 ; delta<-2 ; S<-10000 ; THETA<-NULL ; set.seed(1)

for(s in 1:S)
{

## simulating our proposal
  theta.star<-rnorm(1,theta,sqrt(delta))

##taking the log of the ratio r
  log.r<-( sum(dnorm(y,theta.star,sqrt(s2),log=TRUE)) +
               dnorm(theta.star,mu,sqrt(t2),log=TRUE) )  -
         ( sum(dnorm(y,theta,sqrt(s2),log=TRUE)) +
               dnorm(theta,mu,sqrt(t2),log=TRUE) )

  if(log(runif(1))<log.r) { theta<-theta.star }

##updating THETA

  THETA<-c(THETA,theta)

}

##two plots: trace of theta and comparing the empirical distribution
##of simulated values to the true posterior

pdf("metropolis_normal.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
```

```
skeep<-seq(10,S,by=10)
plot(skeep,THETA[skeep],type="l",xlab="iteration",ylab=expression(theta))

hist(THETA[-(1:50)],prob=TRUE,main="",xlab=expression(theta),ylab="density")
th<-seq(min(THETA),max(THETA),length=100)
lines(th,dnorm(th,mu.n,sqrt(t2.n)) )
dev.off()
```

### ○ Metropolis-Hastings Algorithm

Recall that a Markov chain is a sequentially generated sequence $\{x^{(1)},,x^{(2)},\ldots\}$ such that the mechanism that generates $x^{(s+1)}$ can depend on the value of $x^{(s)}$ but not on anything that was in the sequence before it. A better way of putting this: for a Markov chain, the future depends on the present and not on the past.

The Gibbs sampler and the Metropolis algorithm are both ways of generating Markov chains that approximate a target probability distribution.

We first consider a simple example where our target probability distribution is $p_o(u, v)$, a bivariate distribution for two random variables $U$ and $V$. In the one-sample normal problem, we would have $U = \theta$, $V = \sigma^2$ and $p_o(u, v) = p(\theta, \sigma^2|y)$.

What does the Gibbs sampler have us do? It has us iteratively sample values of $U$ and $V$ from their conditional distributions. That is,

1. update $U$ : sample $u^{(s+1)} \sim p_o(u \mid v^{(s)})$
2. update $V$ : sample $v^{(s+1)} \sim p_o(v \mid u^{(s+1)})$.

In contrast, Metropolis proposes changes to $X = (U, V)$ and then accepts or rejects those changes based on $p_o$. An alternative way to implement the Metropolis algorithm if to propose and then accept or reject change to one element at a time:

1. update $U$ :
   (a) sample $u^* \sim J_u(u \mid u^{(s)})$

   (b) compute $r = \dfrac{p_o(u^*, v^{(s)})}{p_o(u^{(s)}, v^{(s)})}$

   (c) set $u^{(s+1)}$ equal to $u^*$ or $u^{(s+1)}$ with prob min(1,r) and max(0,1-r).

2. update $V$ : sample $v^{(s+1)} \sim p_o(v \mid u^{(s+1)})$.

   (a) sample $v^* \sim J_u(v \mid v^{(s)})$

   (b) compute $r = \dfrac{p_o(u^{(s+1)}, v^*)}{p_o(u^{(s+1)}, v^{(s)})}$

   (c) set $v^{(s+1)}$ equal to $v^*$ or $v^{(s)}$ with prob min(1,r) and max(0,1-r).

Here, $J_u$ and $J_v$ are separate symmetric proposal distributions for $U$ and $V$.

- The Metropolis algorithm generates proposals from $J_u$ and $J_v$
- It accepts them with some probability min(1,r).
- Similarly, each step of Gibbs can be seen as generating a proposal from a full conditional and then accepting it with probability 1.
- The Metropolis-Hastings (MH) algorithm generalizes both of these approaches by allowing arbitrary proposal distributions.
- The proposal distributions can be symmetric around the current values, full conditionals, or something else entirely. A MH algorithm for approximating $p_o(u, v)$ runs as follows:

1. update $U$ :

   (a) sample $u^* \sim J_u(u \mid u^{(s)}, v^{(s)})$

   (b) compute
   $$r = \frac{p_o(u^*, v^{(s)})}{p_o(u^{(s)}, v^{(s)})} \times \frac{J_u(u^{(s)} \mid u^*, v^{(s)})}{J_u(u^* \mid u^{(s)}, v^{(s)})}$$

   (c) set $u^{(s+1)}$ equal to $u^*$ or $u^{(s+1)}$ with prob min(1,r) and max(0,1-r).

2. update $V$ :

   (a) sample $v^* \sim J_v(u \mid u^{(s+1)}, v^{(s)})$

(b) compute

$$r = \frac{p_o(u^{(s+1)}, v^*)}{p_o(u^{(s+1)}, v^{(s)})} \times \frac{J_u(v^{(s+1)} \mid u^{(s+1)}, v^*)}{J_u(v^* \mid u^{(s+1)}, v^{(s)})}$$

(c) set $v^{(s+1)}$ equal to $v^*$ or $v^{(s+1)}$ with prob min(1,r) and max(0,1-r).

In the above algorithm, the proposal distributions $J_u$ and $J_v$ are not required to be symmetric. The only requirement is that they not depend on $U$ or $V$ values in our sequence previous to the most current values. This requirement ensures that the sequence is a Markov chain.

Doesn't the algorithm above look familiar? Yes, it looks a lot like Metropolis, except the acceptance ratio $r$ contains an extra factor:

- It contains the ratio of the prob of generating the current value from the proposed to the prob of generating the proposed from the current.
- This can be viewed as a correction factor.
- If a value $u^*$ is much more likely to be proposed than the current value $u^{(s)}$ then we must down-weight the probability of accepting $u$.
- Otherwise, such a value $u^*$ will be overrepresented in the chain.

Exercise 1: Show that Metropolis is a special case of MH. Hint: Think about the jumps J.

Exercise 2: Show that Gibbs is a special case of MH. Hint: Show that r = 1.

Note: The MH algorithm can easily be generalized.

**Example 5.10:** Poisson Regression We implement the Metropolis algorithm for a Poisson regression model.

- We have a sample from a population of 52 song sparrows that was studied over the course of a summer and their reproductive activities were recorded.
- In particular, their age and number of new offspring were recorded for each sparrow (Arcese et al., 1992).

- A simple probability model to fit the data would be a Poisson regression where, Y = number of offspring conditional on x = age.

Thus, we assume that

$$Y|\theta_x \sim \text{Poisson}(\theta_x).$$

For stability of the model, we assume that the mean number of offspring $\theta_x$ is a smoother function of age. Thus, we express $\theta_x = \beta_1 + \beta_2 x_+ \beta_3 x^2$.

Remark: This parameterization allows some values of $\theta_x$ to be negative, so as an alternative we reparameterize and model the log-mean of Y, so that

$$\log E(Y|x) = \log \theta_x = \log(\beta_1 + \beta_2 x_+ \beta_3 x^2)$$

which implies that

$$\theta_x = \exp(\beta_1 + \beta_2 x_+ \beta_3 x^2) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}).$$

Now back to the problem of implementing Metropolis. For this problem, we will write

$$\log E(Y_i|x_i) = \log(\beta_1 + \beta_2 x_i + \beta_3 x_i^2) = \boldsymbol{\beta}^T \boldsymbol{x_i},$$

where $x_i$ is the age of sparrow $i$. We will abuse notation slightly and write $\boldsymbol{x_i} = (1, x_i, x_i^2)$.

- We will assume the prior on the regression coefficients is iid Normal(0,100).
- Given a current value $\boldsymbol{\beta^{(s)}}$ and a value $\boldsymbol{\beta^*}$ generated from $J(\boldsymbol{\beta^*}, \boldsymbol{\beta^{(s)}})$ the acceptance ration for the Metropolis algorithm is:

$$r = \frac{p(\boldsymbol{\beta^*}|\boldsymbol{X}, \boldsymbol{y})}{p(\boldsymbol{\beta^{(s)}}|\boldsymbol{X}, \boldsymbol{y})} = \frac{\prod_{i=1}^n \text{dpois}(y_i, x_i^T \beta^*)}{\prod_{i=1}^n \text{dpois}(y_i, x_i^T \beta^{(s)})} \times \frac{\prod_{j=1}^3 \text{dnorm}(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \text{dnorm}(\beta_j^{(s)}, 0, 10)}.$$

- We just need to specify the proposal distribution for $\theta^*$
- A convenient choice is a multivariate normal distribution with mean $\boldsymbol{\beta^{(s)}}$.

- In many problems, the posterior variance can be an efficient choice of a proposal variance. But we don't know it here.
- However, it's often sufficient to use a rough approximation. In a normal regression problem, the posterior variance will be close to $\sigma^2(X^TX)^{-1}$ where $\sigma^2$ is the variance of $Y$.

In our problem: $E\log Y = \beta^T x$ so we can try a proposal variance of $\hat{\sigma}^2(X^TX)^{-1}$ where $\hat{\sigma}^2$ is the sample variance of $\log(y+1/2)$.

Remark: Note we add 1/2 because otherwise $\log 0$ is undefined. The code of implementing the algorithm is included below.

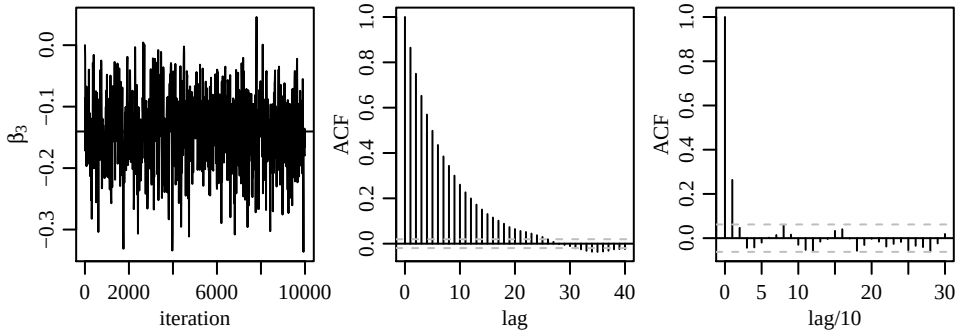



FIGURE 5.15: Plot of the Markov chain in $\beta_3$ along with autocorrelations functions

```
###example 5.10 -- sparrow Poisson regression
yX.sparrow<-dget("http://www.stat.washington.edu/~hoff/Book/Data/data/yX.sparrow")

### sample from the multivariate normal distribution
rmvnorm<-function(n,mu,Sigma)
{
  p<-length(mu)
  res<-matrix(0,nrow=n,ncol=p)
  if( n>0 & p>0 )
  {
    E<-matrix(rnorm(n*p),n,p)
    res<-t(  t(E%*%chol(Sigma)) +c(mu))
  }
```

```
  res
}


y<- yX.sparrow[,1]; X<- yX.sparrow[,-1]
n<-length(y) ; p<-dim(X)[2]

pmn.beta<-rep(0,p)
psd.beta<-rep(10,p)

var.prop<- var(log(y+1/2))*solve( t(X)%*%X )
beta<-rep(0,p)
S<-10000
BETA<-matrix(0,nrow=S,ncol=p)
ac<-0
set.seed(1)

for(s in 1:S) {

#propose a new beta

beta.p<- t(rmvnorm(1, beta, var.prop ))

lhr<- sum(dpois(y,exp(X%*%beta.p),log=T)) -
      sum(dpois(y,exp(X%*%beta),log=T)) +
      sum(dnorm(beta.p,pmn.beta,psd.beta,log=T)) -
      sum(dnorm(beta,pmn.beta,psd.beta,log=T))

if( log(runif(1))< lhr ) { beta<-beta.p ; ac<-ac+1 }

BETA[s,]<-beta
                    }
cat(ac/S,"\n")

#######

library(coda)
apply(BETA,2,effectiveSize)
```

```
####
pdf("sparrow_plot1.pdf",family="Times",height=1.75,width=5)
par(mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
par(mfrow=c(1,3))
blabs<-c(expression(beta[1]),expression(beta[2]),expression(beta[3]))
thin<-c(1,(1:1000)*(S/1000))
j<-3
plot(thin,BETA[thin,j],type="l",xlab="iteration",ylab=blabs[j])
abline(h=mean(BETA[,j]) )

acf(BETA[,j],ci.col="gray",xlab="lag")
acf(BETA[thin,j],xlab="lag/10",ci.col="gray")
dev.off()
####
```

## ○ Metropolis and Gibbs Combined

In complex models, it is often the case that the conditional distributions are available for some parameters but not for others. What can we do then? In these situations we can combine Gibbs and Metropolis-type proposal distributions to generate a Markov chain to approximate the joint posterior distribution of all the parameters.

- Here, we look at an example of estimating the parameters in a regression model for time-series data, where the errors are temporally correlated.
- The full conditionals are available for the regression parameters here, but not the parameter describing the dependence among the observations.

**Example 5.11:** Historical $CO_2$ and temperature data

Analyses of ice cores from East Antarctica have allowed scientists to deduce historical atmospheric conditions of law few hundred years (Petit et al, 1999). Figure 5.18 plots time-series of temperature and carbon dioxide concentration on a standardized scale (centered and called to have mean of zero and variance of 1).

- The data include 200 values of temperature measured at roughly equal time intervals, with time between consecutive measurements being around 2,000 years.
- For each value of temperature there is a $CO_2$ concentration value that corresponds to data that is about 1,000 years previous to the temperature value (on average).
- Temperature is recorded in terms of its difference from current present temperature in degrees Celsius and $CO_2$ concentration is recorded in parts per million by volume.

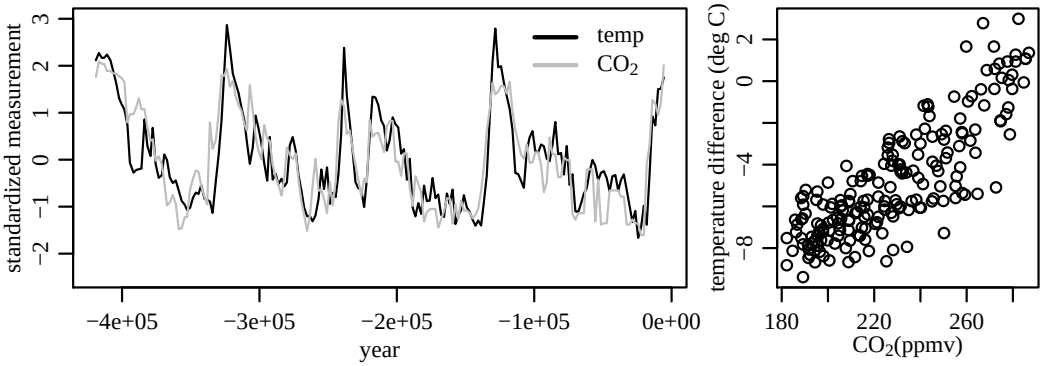



FIGURE 5.16: Temperature and carbon dioxide data.

- The plot indicates the temporal history of temperature and $CO_2$ follow very similar patterns.
- The second plot in Figure 5.18 indicates that $CO_2$ concentration at a given time is predictive of temperature following that time point.
- We can quantify this using a linear regression model for temperature $(Y)$ as a function of $(CO_2)(x)$.
- The validity of the standard error relies on the error terms in the regression model being iid and standard confidence intervals further rely on the errors being normally distributed.
- These two assumptions are examined in the two residual diagnostic plots in Figure 5.19.

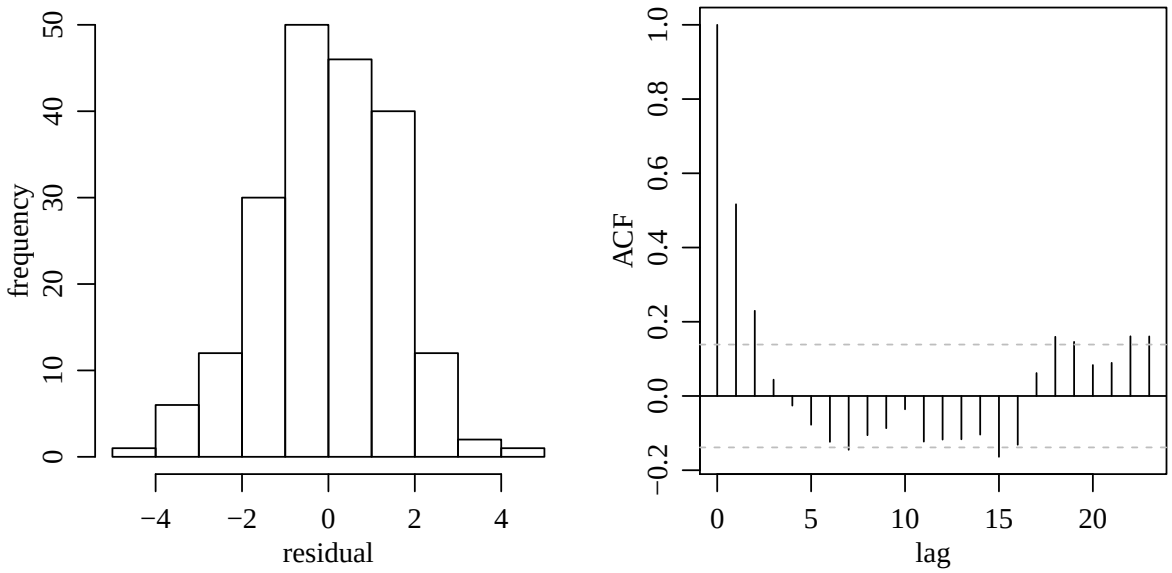



FIGURE 5.17: Temperature and carbon dioxide data.

- The first plot shows a histogram of the residuals and indicates no serious deviation from non-normality.
- The second plot gives the autocorrelation function of the residuals, indicating a nontrivial correlation of 0.52 between residuals at consecutive time points.
- Such a positive correlation generally implies there is less information in the data and less evidence for a relations between the two variables than is assumed by the OLS regression analysis.

The ordinary regression model is

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}).$$

The diagnostic plots suggest that a more appropriate model for the ice core data is one in which the error terms are not independent, but temporally correlated.

We will replace $\sigma^2 \boldsymbol{I}$ with a covariance matrix $\Sigma$ that can represent the positive correlation between sequential observations. One simple, popular class of covariance matrices for temporally correlated data are those having *first order autoregressive structure*:

$$\Sigma = \sigma^2 C_p = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \\ & & & & \\ \vdots & \vdots & & \ddots & \\ & & & & \\ \rho^{n-1} & \rho^{n-2} & & & 1 \end{pmatrix}$$

Under this covariance matrix the variance of $Y_i|\beta, \boldsymbol{x}_i$ is $\sigma^2$ but the correlation between $Y_i$ and $Y_{i+t}$ is $\rho^t$. Using the multivariate normal and inverse gamma prior (it is left as an exercise to show that)

$$\begin{aligned} \boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2, \rho &\sim N(\beta_n, \Sigma_n), \\ \sigma^2 \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta} &\sim IG((\nu_o + n)/2, [\nu_o \sigma_o^2 + SSR_\rho]/2) \end{aligned}$$

where $\beta_n = \Sigma_n \left( \boldsymbol{X}^T C_p^{-1} \boldsymbol{X}/\sigma^2 + \Sigma_o^{-1} \beta_o \right)^{-1}$ and $\Sigma_n = \left( \boldsymbol{X}^T C_p^{-1} \boldsymbol{X}/\sigma^2 + \Sigma_o^{-1} \right)^{-1}$ and $SSR_\rho = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T C_p^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$

- If $\beta_o$ and $\Sigma_o$ has large diagonal entries, then $\beta_n$ is very close to

$$(\boldsymbol{X}^T C_p^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T C_p^{-1} \boldsymbol{y}$$

- If $\rho$ were known this would be the generalized least squares (GLS) estimate of $\beta$.

- This is a type of weighted LS estimate that is used when the error terms are not iid. In such situations, both OLS and GLS provide unbiased estimates of $\beta$ but the GLS has lower variance.
- Bayesian analysis using a model that accounts for correlation errors provides parameter estimates that are similar to those of GLS, so for convenience we will refer to our analysis as "Bayesian GLS."

If we knew the value of $\rho$ we could just implement Gibbs to approximate $p(\beta, \sigma^2|X, y, \rho)$. However, $\rho$ is unknown and typically the distribution of $\rho$ is nonstandard for most prior distributions, suggesting that the Gibbs sampler isn't applicable. What can we do instead?

We can use the generality of the MH algorithm. Recall we are allowed to use different proposals at each step. We can iteratively update $\beta, \sigma^2$, and $\rho$ at different steps (using Gibbs proposals). That is:

- We will make proposals for $\beta$ and $\sigma^2$ using the full conditionals and
- make a symmetric proposal for $\rho$.
- Following the rules of MH, we accept with prob 1 any proposal coming from a full conditional distribution, whereas we have to calcite an acceptance probability for proposals of $\rho$.

We run the following algorithm:

1. Update $\beta$: Sample $\beta^{(s+1)} \sim N(\beta_n, \Sigma_n)$, where $\beta_n$ and $\Sigma_n$ depend on $\sigma^{2(s)}$ and $\rho^{(s)}$.

2. Update $\sigma^2$: Sample $\sigma^{2(s+1)}$ ~IG( $(\nu_o + n)/2, [\nu_o\sigma_o^2 + SSR_\rho]/2$) where $SSR_\rho$ depends on $\beta^{(s+1)}$ and $\rho^{(s)}$.

3. Update $\rho$ : (a): Propose $\rho^* \sim \text{Uniform}(\rho^{(s)} - \delta, \rho^{(s)} + \delta)$. If $\rho^* < 0$ then reassign it to be $|p^*|$. If $\rho^* > 1$ then reassign it to be $2 - \rho^*$.

   (b) Compute the acceptance ratio

$$r = \frac{p(y \mid X, \beta^{(s+1)}, \sigma^{2(s+1)}, \rho^*)\, p(\rho^*)}{p(y \mid X, \beta^{(s+1)}, \sigma^{2(s+1)}, \rho^{(s)})\, p(\rho^{(s)})} andsample$$

   u ~ Uniform$(0, 1)$. If $u < r$, set $\rho^{(s+1)} = \rho^*$, otherwise $\rho^{(s+1)} = \rho^{(s)}$.

The proposal used in Step 3(a) is called *reflecting random walk*, which insures that $0 < \rho < 1$. Note that a sequence of MH steps in which each parameter is updated is often referred to as a *scan* of the algorithm.

Exercise: Show that the proposal is symmetric.

For convenience and ease, we're going to use diffuse priors for the parameters with $\beta_o = 0, \Sigma_o = diag(1000), \nu_o = 1$, and $\sigma^2 = 1$. Our prior on $\rho$ will be Uniform(0,1). We first run 1000 iterations of the MH algorithm and show a trace plot of $\rho$ as well as an autocorrelation plot (Figure 5.20).

Suppose now we want to generate 25,000 scans for a total of 100,000 parameter values. The MC is highly correlated, so we will thin every 25th value in the chain. This reduces the autocorrelation.

The Monte Carlo approximation of the posterior density of $\beta_2$ (the slope) appears in the Figure 5.20. The posterior mean is 0.028 with 95 percent posterior credible interval of (0.01,0.05), indicating that the relationship between temperature and $CO_2$ is positive. As indicated in the second plot this relationship seems much weaker than suggested by the OLS estimate of 0.08. For the OLS estimation, the small number of data points with high y-values have a large influence on the estimate of $\beta$. On the other hand, the GLS model recognizes many of these extreme points are highly correlated with one another and down weights their influence.

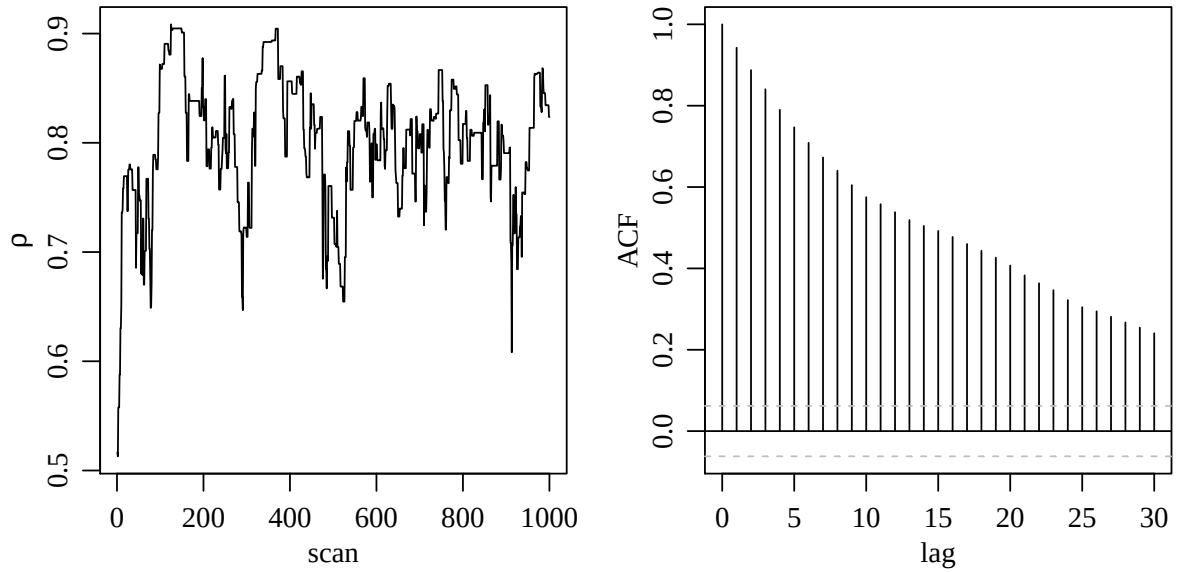



FIGURE 5.18: The first 1,000 values of $\rho$ generated from the Markov chain.

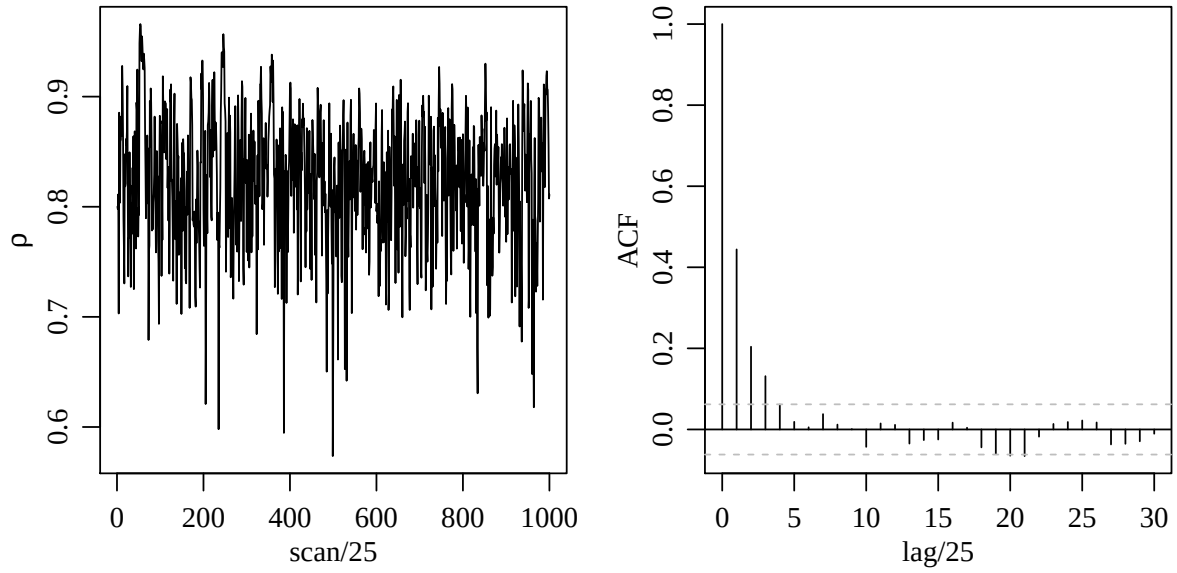



FIGURE 5.19: Every 25th value of $\rho$ generated from the Markov chain of length 25,000.

Remark: this weaker regression coefficient is a result of the temporally correlated data and not of the particular prior distribution we used or the Bayesian approach in general.

Exercise: Repeat the analysis with different prior distributions and perform non-Bayesian GLS for comparison.

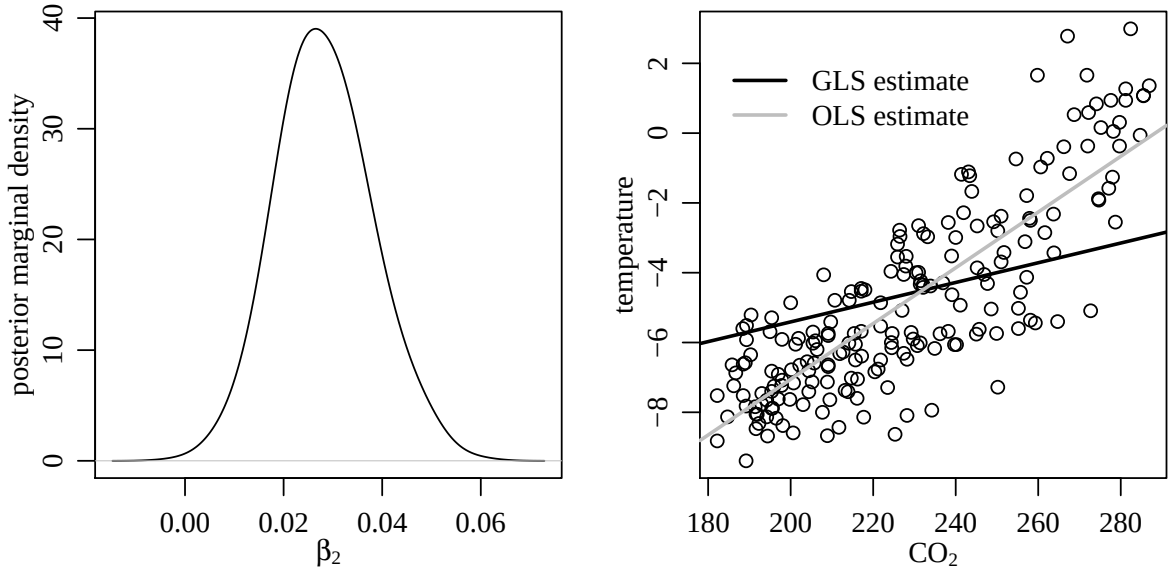



FIGURE 5.20: Posterior distribution of the slope parameter $\beta_2$ and posterior mean regression line (after generating the Markov chain with length 25,000 with thin 25).

```
#####
##example 5.10 in notes
# MH and Gibbs problem
##temperature and co2 problem

source("http://www.stat.washington.edu/~hoff/Book/Data/data/chapter10.r")


### sample from the multivariate normal distribution
rmvnorm<-function(n,mu,Sigma)
{
```

```
  p<-length(mu)
  res<-matrix(0,nrow=n,ncol=p)
  if( n>0 & p>0 )
  {
    E<-matrix(rnorm(n*p),n,p)
    res<-t(  t(E%*%chol(Sigma)) +c(mu))
  }
  res
}
###

##reading in the data and storing it
dtmp<-as.matrix(read.table("volstok.txt",header=F), sep = "-")
dco2<-as.matrix(read.table("co2.txt",header=F, sep = "\t"))
dtmp[,2]<- -dtmp[,2]
dco2[,2]<- -dco2[,2]
library(nlme)

#### get evenly spaced temperature points
ymin<-max( c(min(dtmp[,2]),min(dco2[,2])))
ymax<-min( c(max(dtmp[,2]),max(dco2[,2])))
n<-200
syear<-seq(ymin,ymax,length=n)
dat<-NULL
for(i in 1:n) {
 tmp<-dtmp[ dtmp[,2]>=syear[i] ,]
 dat<-rbind(dat,  tmp[dim(tmp)[1],c(2,4)] )
               }
dat<-as.matrix(dat)
####

####
dct<-NULL
for(i in 1:n) {
  xc<-dco2[ dco2[,2] < dat[i,1] ,,drop=FALSE]
  xc<-xc[ 1, ]
  dct<-rbind(dct, c( xc[c(2,4)], dat[i,] ) )
               }

mean( dct[,3]-dct[,1])
```

```
dct<-dct[,c(3,2,4)]
colnames(dct)<-c("year","co2","tmp")
rownames(dct)<-NULL
dct<-as.data.frame(dct)

##looking at temporal history of co2 and temperature
########
pdf("temp_co2.pdf",family="Times",height=1.75,width=5)
par(mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
layout(matrix( c(1,1,2),nrow=1,ncol=3) )

#plot(dct[,1],qnorm( rank(dct[,3])/(length(dct[,3])+1 )) ,
plot(dct[,1],  (dct[,3]-mean(dct[,3]))/sd(dct[,3]) ,
   type="l",col="black",
   xlab="year",ylab="standardized measurement",ylim=c(-2.5,3))
legend(-115000,3.2,legend=c("temp",expression(CO[2])),bty="n",
       lwd=c(2,2),col=c("black","gray"))
lines(dct[,1],  (dct[,2]-mean(dct[,2]))/sd(dct[,2]),
#lines(dct[,1],qnorm( rank(dct[,2])/(length(dct[,2])+1 )),
  type="l",col="gray")

plot(dct[,2], dct[,3],xlab=expression(paste(CO[2],"(ppmv)")),
ylab="temperature difference (deg C)")
dev.off()
########


##residual analysis for the least squares estimation
########
pdf("residual_analysis.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))

lmfit<-lm(dct$tmp~dct$co2)
hist(lmfit$res,main="",xlab="residual",ylab="frequency")
#plot(dct$year, lmfit$res,xlab="year",ylab="residual",type="l" ); abline(h=0)
acf(lmfit$res,ci.col="gray",xlab="lag")
dev.off()
```

```
########

##BEGINNING THE GIBBS WITHIN METROPOLIS

######## starting values (DIFFUSE)
n<-dim(dct)[1]
y<-dct[,3]
X<-cbind(rep(1,n),dct[,2])
DY<-abs(outer( (1:n),(1:n) ,"-"))

lmfit<-lm(y~-1+X)
fit.gls <- gls(y~X[,2], correlation=corARMA(p=1), method="ML")
beta<-lmfit$coef
s2<-summary(lmfit)$sigma^2
phi<-acf(lmfit$res,plot=FALSE)$acf[2]
nu0<-1 ; s20<-1 ; T0<-diag(1/1000,nrow=2)
###
set.seed(1)

###number of MH steps
S<-25000 ; odens<-S/1000
OUT<-NULL ; ac<-0 ; par(mfrow=c(1,2))
library(psych)
for(s in 1:S)
{

  Cor<-phi^DY  ; iCor<-solve(Cor)
  V.beta<- solve( t(X)%*%iCor%*%X/s2 + T0)
  E.beta<- V.beta%*%( t(X)%*%iCor%*%y/s2  )
  beta<-t(rmvnorm(1,E.beta,V.beta)  )

  s2<-1/rgamma(1,(nu0+n)/2,(nu0*s20+t(y-X%*%beta)%*%iCor%*%(y-X%*%beta)) /2 )

  phi.p<-abs(runif(1,phi-.1,phi+.1))
  phi.p<- min( phi.p, 2-phi.p)
  lr<- -.5*( determinant(phi.p^DY,log=TRUE)$mod -
             determinant(phi^DY,log=TRUE)$mod  +
   tr( (y-X%*%beta)%*%t(y-X%*%beta)%*%(solve(phi.p^DY) -solve(phi^DY)) )/s2 )

  if( log(runif(1)) < lr ) { phi<-phi.p ; ac<-ac+1 }
```

```
  if(s%%odens==0)
    {
      cat(s,ac/s,beta,s2,phi,"\n") ; OUT<-rbind(OUT,c(beta,s2,phi))
#      par(mfrow=c(2,2))
#      plot(OUT[,1]) ; abline(h=fit.gls$coef[1])
#      plot(OUT[,2]) ; abline(h=fit.gls$coef[2])
#      plot(OUT[,3]) ; abline(h=fit.gls$sigma^2)
#      plot(OUT[,4]) ; abline(h=.8284)

    }
}
#####

OUT.25000<-OUT
library(coda)
apply(OUT,2,effectiveSize )


OUT.25000<-dget("data.f10_10.f10_11")
apply(OUT.25000,2,effectiveSize )


pdf("trace_auto_1000.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
plot(OUT.1000[,4],xlab="scan",ylab=expression(rho),type="l")
acf(OUT.1000[,4],ci.col="gray",xlab="lag")
dev.off()


pdf("trace_thin_25.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
plot(OUT.25000[,4],xlab="scan/25",ylab=expression(rho),type="l")
acf(OUT.25000[,4],ci.col="gray",xlab="lag/25")
dev.off()

pdf("fig10_11.pdf",family="Times",height=3.5,width=7)
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
```

```
par(mfrow=c(1,2))

plot(density(OUT.25000[,2],adj=2),xlab=expression(beta[2]),
   ylab="posterior marginal density",main="")

plot(y~X[,2],xlab=expression(CO[2]),ylab="temperature")
abline(mean(OUT.25000[,1]),mean(OUT.25000[,2]),lwd=2)
abline(lmfit$coef,col="gray",lwd=2)
legend(180,2.5,legend=c("GLS estimate","OLS estimate"),bty="n",
      lwd=c(2,2),col=c("black","gray"))
dev.off()

quantile(OUT.25000[,2],probs=c(.025,.975) )

plot(X[,2],y,type="l")
points(X[,2],y,cex=2,pch=19)
points(X[,2],y,cex=1.9,pch=19,col="white")
text(X[,2],y,1:n)

iC<-solve( mean(OUT[,4])^DY )
Lev.gls<-solve(t(X)%*%iC%*%X)%*%t(X)%*%iC
Lev.ols<-solve(t(X)%*%X)%*%t(X)

plot(y,Lev.ols[2,] )
plot(y,Lev.gls[2,] )
```

## 5.6 Introduction to Nonparametric Bayes

As we have seen, Bayesian parametric methods takes classical methodology for prior and posterior distributions in models with a finite number of parameters. It is often the case that the number of parameters taken in such model is low for computational complexity, however, in current research problem we deal with high dimensional data and high dimensional parameters. The origins of Bayesian methods have been around since the mid-1700's and are still thriving today. The applicability of Bayesian parametric models still remains and has widened with the increased advancements made in modern computing and the growth of methods available in Markov chain Monte Carlo.

Frequentist nonparametrics covers a wide array of areas in statistics. The area is well known for being associated with testing procedures that are or become asymptotically distribution free, which lead to nonparametric confidence intervals, bands, etc. (Hjors et al., 2010). Further information can be found on these methods in Wasserman (2006).

Nonparametric Bayesian methods are models and methods characterized generally by large parameter spaces, such as unknown density and regression functions and construction of probability measures over these spaces. Typical examples seen in practice include density estimation, nonparametric regression with fixed error distributions, hazard rate and survival function estimation. For a thorough introduction into this subject see (Hjors et al., 2010).

## ○ Motivations

The motivation is the following:

- We have $X_1 \dots X_n \overset{iid}{\sim} F, F \in \mathcal{F}$. We usually assume that $\mathcal{F}$ is a parametric family.
- Then, putting a prior on $\mathcal{F}$ amounts to putting a prior on $\mathbb{R}^d$ for some $d$.
- We would like to be able to put a prior on all the set of cdf's. And we would like the prior to have some basic features:

1. The prior should have large support.
2. The prior should give rise to priors which are analytically tractable or computationally manageable.
3. We should be able to center the prior around a given parametric family.

## ○ The Dirichlet Process

### Review of Finite Dimensional Dirichlet Distribution

- This is a distribution on the k-dimensional simplex.
- Let $(\alpha_1, \ldots, \alpha_k)$ be such that $\alpha_j > 0$ for all $j$.
- The Dirichlet distribution with parameter vector $(\alpha_1, \ldots, \alpha_k)$ has density
$$p(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1}.$$
- It is conjugate to the Multinomial distribution. That is if $\boldsymbol{Y} \sim Multinomial(n, \boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim Dir(\alpha_1, \ldots, \alpha_k)$, then it can be shown that
$$\boldsymbol{\theta} | \boldsymbol{y} \sim Dir(\alpha_1 + N_1, \ldots, \alpha_k + N_k).$$
- It can be shown that
$$E(\theta_j) = \alpha_j / \alpha$$
where $\alpha = \sum_j \alpha_j$. It can also be shown that
$$Var(\theta_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}.$$

**Infinite Dimensional Dirichlet Distribution**

Let $\alpha$ be a finite (non-null) measure (or think probability distribution) on $\mathbb{R}$. Sometimes $\alpha$ will be called the concentration parameter (in scenarios when we might hand wave the measure theory for example or it's not needed).

You should think about the Infinite Dimension Dirichlet Distribution as a distribution of distributions as we will soon see.

DEFINITION 5.1: $F$ has the Dirichlet distribution with parameter (measure) $\alpha$ if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathbb{R}$ the $k$-dimensional random vector $(F(\{A_1\}), \ldots, F(\{A_k\}))$ has the finite k-dimensional Dirichlet distribution

$$\text{Dir}(\alpha(A_1), \ldots, \alpha(A_k)).$$

For more on this see: Freedman (1963), Ferguson (1973, 1974).

**Intuition**: Each $F(\{A_k\}) \in [0, 1]$ since $F$ is some cumulative distribution function. Also,

$$F(\{A_1\}) + \cdots + F(\{A_k\}),$$

thus, $(F(\{A_1\}), \ldots, F(\{A_k\}))$ lives on the $k$-dimensional simplex.

**Remark**: For those with measure theory: You can't have a measure that is 0. Note that Lebesgue measure isn't finite on the reals.

We will construct the Dirichlet process to intuitively understand it based on the "Polya Urn Scheme" of Blackwell and MacQueen (1973). This is one of the most intuitive approaches. Others in the literature include Ferguson (1973, 1974), which include two constructions. There is an incorrect constructions involve the Kolmovgorov extension theorem (the problem is that the sets aren't measurable, so an important technical detail). The other is a correct construction based on something called the gamma process (this involves much overhead and existence of the gamma process).

### ○ Polya Urn Scheme on Urn With Finitely Many Colors

- Consider an urn containing a finite number of balls of $k$ different colors.
- There are $\alpha_1, \ldots, \alpha_k$ balls of colors $1 \ldots, k$, respectively.
- We pick a ball at random, look at its color, return it to the urn together with another ball of the same color.
- We repeat this indefinitely.
- Let $p_1(n), \ldots p_k(n)$ be the proportions of balls of colors $1 \ldots, k$ at time $n$.

**Example 5.12:** Polya Urn for Three Balls
Suppose we have three balls in our urn. Let red correspond the ball 1. Let blue correspond the ball 2. Let green correspond the ball 3. Furthermore, suppose that $P(red) = 2/9), P(blue) = 3/9$ and $P(green) = 4/9$.

Let $\alpha$ be a the following probability measure (or rather discrete probability distribution):

- $\alpha_o(1) = 2/9$.
- $\alpha_o(2) = 3/9$.
- $\alpha_o(3) = 4/9$.

Another way of writing this is define $\alpha_o = 2/9\ \delta_1 + 3/9\ \delta_2 + 4/9\ \delta_3$ where

$$\delta_1(A) = \begin{cases} 1 & \text{if } 1 \in A \\ 0 & \text{otherwise.} \end{cases}$$

### ○ Polya Urn Scheme in General

Let $\alpha$ be a finite measure on a space $\mathcal{X} = \mathbb{R}$.

1. (Step 1) Suppose $X_1 \sim \alpha_o$.
2. (Step 2) Now create a new measure $\alpha + \delta_{X_1}$ where

$$\delta_{X_1}(A) = \begin{cases} 1 & \text{if } X_1 \in A \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$X_2 \sim \frac{\alpha + \delta_{X_1}}{\alpha() + \delta_{X_1}()} = \frac{\alpha + \delta_{X_1}}{\alpha() + 1}.$$

Fact: $\delta_{X_1}() = 1$. Think about why this is intuitively true.

**What does the above equation really mean?**

- $\alpha$ represents the original distribution of balls.
- $\delta_{X_1}$ represents the ball we just added.

**Deeper understanding**

- Suppose the urn contained $N$ total balls when we started.
- Then the probability that the second ball drawn $X_2$ will be of the original $N$ balls is $N/(N+1)$.
- This represents the $\alpha$ part of the distribution of $X_2$.
- We want the probability of drawing a new ball to be $1/(N+1)$. This goes with $\delta_{X_1}$.
- When we write $X_2 \sim \frac{\alpha + \delta_{X_1}}{\text{norm. constant}}$ we want $N/(N+1)$ of the probability to go to $\frac{\alpha}{\text{norm. constant}}$ and $1/(N+1)$ to go to $\frac{\delta_{X_1}}{\text{norm. constant}}$.

How does this continue? Since we want

$$= \frac{\delta_{X_1}()}{\alpha() + 1} = 1/(N+1)$$

this implies that

$$\frac{1}{\alpha() + 1} = 1/(N+1) \implies \alpha() = N.$$

Hence, we take $\alpha() = N$ and then we plug back in and find that $\alpha_o = \alpha/N \implies \alpha = \alpha_o N$.

This implies that

$$X_2 \sim \frac{\alpha_o N + \delta_{X_1}}{N+1},$$

which is now in terms of $\alpha_o$ and $N$ (which we know).

(Step 3) Continue forming new measures: $\alpha + \delta_{X_1} + \delta_{X_2}$. Then

$$X_3 \sim \frac{\alpha + \delta_{X_1} + \delta_{X_2}}{\alpha() + \delta_{X_1}() + \delta_{X_2}()} = \frac{\alpha + \delta_{X_1} + \delta_{X_2}}{\alpha() + 2}.$$

In general, it can be shown that

$$P(X_{n+1} \mid X_1 \ldots X_n) = \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{X_i}(A)}{\alpha() + n} = \frac{\alpha_o N + \sum_{i=1}^{n} \delta_{X_i}(A)}{N + n}.$$

**Polya Urn Scheme in General Case: Theorem**

- Let $\alpha$ be a finite measure on a space $X$ (this space can be very general, but we will assume it's the reals).
- Define a sequence $\{X_1, X_2, \ldots\}$ of random variables to be a *Polya urn sequence with parameter measure* $\alpha$ if
  - $P(X_1 \in B) = \alpha(B)/\alpha()$.
  - For every $n$,

$$P(X_{n+1} \in B \mid X_1 \ldots, X_n) = \frac{\alpha(B) + \sum_i \delta_{X_i}(B)}{\alpha() + n}.$$

Specifically, $X_1, X_2, \ldots,$ is PUS($\alpha$) if

$$P(X_1 \in A) = \frac{\alpha(A)}{\alpha()} = \alpha_o$$

for every $A \in$ and for every $n$

$$P(X_{n+1} \in B \mid X_1 \ldots, X_n) = \frac{\alpha(B) + \sum_i \delta_{X_i}(B)}{\alpha() + n}$$

for every $A \in$ .

### ○ De Finetti and Exchaneability

Recall what exchangeability means. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a sequence of random variables. This sequence is said to be exchangeable if the distribution of

$$(Y_1, Y_2, \ldots, Y_n) \stackrel{d}{=} (Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(n)})$$

for every permutation $\pi$ of $1, \ldots, n$.

Note: This means that we can permute the random variables and the distribution doesn't change.

An infinite sequence is said to be exchangeable if for every $n$, $Y_1, Y_2, \ldots, Y_n$ is exchangeable. That is, we don't require exchangeability for infinite permutations, but it must be true for every "chunk" that we take that is of length or size $n$.

Definition 5.2: De Finetti's General Theorem
Let $X_1, X_2 \ldots$ be an infinite exchangeable sequence of random variables. Then there exists a probability measure $\pi$ such that

$$X_1, X_2 \ldots, | \, F \overset{iid}{\sim} F$$

$$F \sim \pi$$

for any $x_1, \ldots x_n \in \{0, 1\}$.

Remark: Suppose that $X_1, X_2 \ldots$ is an infinite exchangeable sequence of binary random variables. Then there exists a probability measure (distribution) on $[0, 1]$ such that for every $n$

$$P(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 p^{\Sigma_i x_i} (1-p)^{n - \Sigma_i x_i} \mu(p) dp$$

where $\mu(p)$ is the measure or probability distribution or prior that we take on $p$.

**Theorem 5.1.** *A General Result (without proof)*
*Let $X_1, X_2 \ldots$ be PUS($\alpha$). Then this can be thought of as a two-stage process where*

- $F \sim Dir(\alpha)$
- $X_1, X_2 \ldots, \mid F \overset{iid}{\sim} F$

*If we consider the process of the PUS consisting of $X_2, X_3 \ldots$, then it's a PUS($\alpha + \delta_{X1}$). That is, $F \mid X_1 \sim Dir(\alpha + \delta_{X1})$.*

*More generally, it can be shown that*

$F \mid X_1 \ldots X_n \sim Dir(\alpha + \sum_{i=1}^{n} \delta_{Xi})$.

### ○ Chinese Restaurant Process

- There are Bayesian NP approaches to many of the main issues in statistics including:
  - regression.
  - classification.
  - clustering.
  - survival analysis.
  - time series analysis.
  - spatial data analysis.
- These generally involve assumptions of exchangeability or partial exchangeability.
  - and corresponding distributions on random objects of various kinds (functions, partitions, measures, etc.)
- We look at the problem of clustering for concreteness.

### ○ Clustering: How to choose $K$?

- Adhoc approaches (hierarchical clustering)
  - these methods do yield a data-drive choice of $K$
  - there is little understanding how good these choices are (meaning the checks are adhoc based on some criterion)
- Methods based on objective functions ($M$-estimators)
  - $K$-means, spectral clustering
  - they come with some frequentist guarantees
  - it's often hard to turn these into data-driven choices of $K$
- Parametric likelihood-based methods
  - finite mixture models, Bayesian variants
  - various model choice methods: hypothesis testing, cross-validation, bootstrap, AIC, BIC, Laplace, reversible jump MCMC
  - do the assumptions underlying the method apply to the setting (not very often)
- Something different: The Chinese restaurant process.

Basic idea: In many data analysis settings, we don't know the number of latent clusters and would like to learn it from the data. BNP clustering addresses this by assuming there is an infinite number of latent clusters, but that only a finite number of them is used to generate the observed data. Under these assumptions, the posterior yields a distribution over the number of clusters, the assign of data to clusters, and the parameters associated with each cluster. In addition, the predictive distribution, the assignment of the next data point, allows for new data to be assign to a previously unseen cluster.

How does it work: The BNP problem addresses and finesses the clustering problem by choosing the number of clusters by assuming it is infinite, however it specifies a prior over the infinite groupings $P(c)$ in such a way that favors assigning data to a small number of groups, where $c$ refers to the cluster assignments. The prior over groupings is a well known problem called the Chinese restaurant process (CRP), which is a distribution over infinite partition of the integers (Aldous, 1985; Pitman, 2002).

Where does the name come from?

- Imagine that Sam and Mike own a restaurant with an infinite number of tables.
- Imagine a sequence of customers entering their restaurant and sitting down.
- The first customer (Liz) enters and sits at the first table.
- The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$ and a new table with probability $\frac{\alpha}{1+\alpha}$, where $\alpha$ is positive and real.
- Liz is friendly and people would want to sit and talk with her. So, we would assume that $\frac{1}{1+\alpha}$ is a high probability, meaning that $\alpha$ is a small number.
- What happens with the $n$th customer?
  - He sits at each of the previously occupied tables with probability proportional to the number previous customers sitting there.
  - He sits at the next unoccupied table with probability proportional to $\alpha$.

More formally, let $c_n$ be the table assigned me of customer $n$. A draw from this distribution can be generated by sequentially assigning observations with probability

$$P(c_n = k \mid c) = \begin{cases} \frac{m_k}{n-1+\alpha} & \text{if } k \leq K_+ \text{ (i.e. k is a previously occupied table),} \\ \frac{\alpha}{n-1+\alpha} & \text{otherwise (i.e. k is the next unoccupied table),} \end{cases}$$

where $m_k$ is the number of customers sitting at table $k$ and $K_+$ is the number of table for which $m_k > 0$. The parameter $\alpha$ is called the concentration parameter.

**The rich just get richer**

- CRP rule: next customer sits at a table with prob. proportional to number of customers already sitting at it (and sits at new table with prob. proportional to $\alpha$).
- Customers tend to sit at most popular tables.
- Most popular tables attract the most new customers, and become even more popular.
- CRPs exhibit power law behavior, where a few tables attract the bulk of the customers.
- The concentration parameter $\alpha$ determines how likely a customer is to sit at a fresh table.

More formally stated:

- A larger value of $\alpha$ will produce more occupied tables (and fewer customers per table).
- Thus, a small value of $\alpha$ produces more customers at each table.
- The CRP exhibits an important invariance property: the cluster assignments under this distribution are exchangeable.
- This means that $p(c)$ is unchanged if the order of customers is shuffled (up to label changes). This may be counter-intuitive since the process was just described sequentially.

**The CRP and Clustering**

- The data points refer to the customers and the tables are the clusters.
  - Then the CRP defines a prior distribution on the partition of the data and on the number of tables.
- The prior can be completed with:
  - A likelihood, meaning there needs to be an parameterized probability distribution that corresponds to each table
  - A prior for the parameters –the first customer to sit at table $k$ chooses the parameter vector for that table ($\phi_k$) from the prior
- Now that we have a distribution for any quantity we care about in some clustering setting.

Now, let's think about how we would write down this process out formally. We're writing out a mixture model with a component that's nonparametric.

Let's define the following:

- $y_n$ are the observations at time $n$.
- $c_n$ are the latent clusters that generate $c_n$.
- $F$ is a parametric family of distributions for $y_n$.
- $\theta_k$ represent the clustering parameters.
- $G_o$ represents a general prior for the clustering parameters (this is the nonparametric part).

We also assume that each observation is conditionally independent given its latent cluster assignment and its cluster parameters.

Using the CRP, we can view the model as

$$\begin{aligned} y_n \mid c_n, \theta &\sim F(\theta_{c_n}) \\ c_n &\propto p(c_n) \\ \theta_k &\propto G_o. \end{aligned}$$

We want to know $p(y \mid c)$.

Then by Bayes' rule, $p(c|y) = \frac{p(y|c)p(c)}{\sum_c p(y|c)p(c)}$, where

$$p(y \mid c) = \int_\theta \left[ \prod_{n=1}^{N} F(y|\theta_{c_n}) \prod_{k=1}^{K} G_o(\theta_k) \right] d\theta.$$

A $G_o$ that is conjugate allow this integral to be calculated analytically. For example, the Gaussian is the conjugate prior to a Gaussian with fixed variance (and thus a mixture of Gaussians model is computationally convenient). We illustrate this specific example below.

**Example 5.13:** Suppose

$$\begin{aligned} y_n \mid c_n, \theta &\sim N(\theta_{c_n}, 1) \\ c_n &\sim \text{Multinomial}(1, p) \\ \theta_k &\sim N(\mu, \tau^2), \end{aligned}$$

where $p, \mu$, and $\tau^2$ are known.

Then

$$p(y|c) = \int_\theta \left[ \prod_{n=1}^{N} \text{Normal}(\theta_{c_n}, 1)(y_n) \times \prod_{k=1}^{K} \text{Normal}(\mu, \tau^2)(\theta_k) \right] d\theta.$$

The term above (inside the integral) is just another normal as a function of $\theta$. Then we can integrate $\theta$ out as we have in problems before.

Once we calculate $p(y|c)$, we can simply plug this and $p(c)$ into

$$p(c|y) = \frac{p(y|c)p(c)}{\sum_c p(y|c)p(c)}.$$

**Example 5.14:** Gaussian Mixture using `R`

Information on the `R` package `profdpm`:

This package facilitates inference at the posterior mode in a class of conjugate product partition models (PPM) by approximating the maximum a posteriori data (MAP) partition. The class of PPMs is motivated by an augmented formulation of the Dirichlet process mixture, which is currently the ONLY available member of this class. The profdpm package consists of two model fittting functions, `profBinary` and `profLinear`, their associated summary methods summary.profBinary

and summary.profLinear, and a function (pci) that computes several metrics of agreement between two data partitions. However, the profdpm package was designed to be extensible to other types of product partition models. For more on this package, see help(profdpm) after installation.

- The following example simulates a dataset consisting of 99 longitudinal measurements on 33 units of observation, or subjects.
- Each subject is measured at three times, drawn uniformly and independently from the unit interval.
- Each of the three measurements per subject are drawn independently from the normal distribution with one of three linear mean functions of time, and with unit variance.
- The linear mean functions vary by intercept and slope. The longitudinal structure imposes a grouping among measurements on a single subject.
- Observations grouped in this way should always cluster together. A grouping structure is specified using the group parameter; a factor that behaves similarly to the groups parameter of lattice graphics functions.
- For the PPM of conjugate binary models, the grouping structure is imposed by the model formula.
- Grouped observations correspond to rows of the model matrix, resulting from a call to model.matrix on the formula passed to profBinary. Hence, the profBinary function does not have a group parameter in its prototype.
- The goal of the following example is to recover the simulated partition and to create simultaneous 95% credible bands for the mean within each cluster. The following R code block creates and the simulated dataset.

```
set.seed(42)
sim <- function(multiplier = 1) {
x <- as.matrix(runif(99))
a <- multiplier * c(5,0,-5)
s <- multiplier * c(-10,0,10)
y <- c(a[1]+s[1]*x[1:33],
```

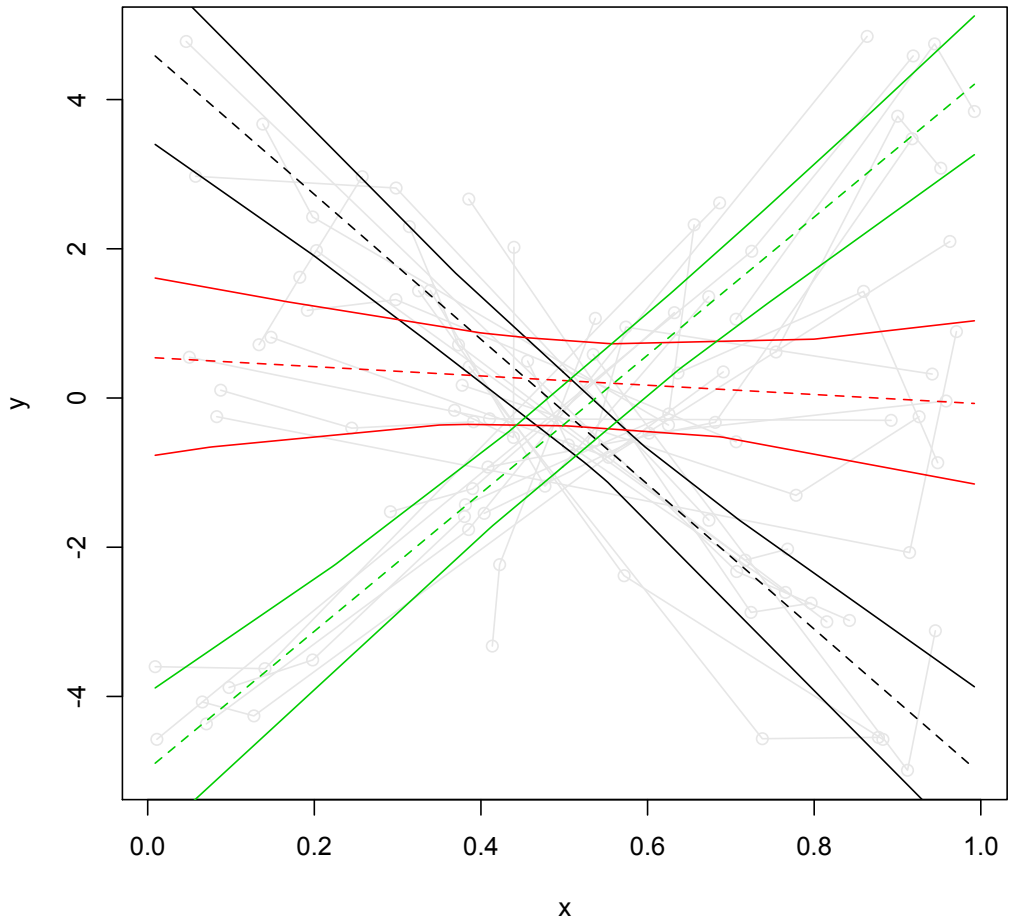



FIGURE 5.21: Simulated data; 99 longitudinal measurements on 33 subjects. Simultaneous confidence bands for the mean within each of the three clusters.

```
a[2]+s[2]*x[34:66],
a[3]+s[3]*x[67:99]) + rnorm(99)
group <- rep(1:33, rep(3,33))
return(data.frame(x=x,y=y,gr=group))
}
dat <- sim()
library("profdpm")
fitL <- profLinear(y ~ x, group=gr, data=dat)
sfitL <- summary(fitL)
%pdf(np_plot.pdf)
plot(fitL$x[,2], fitL$y, col=grey(0.9), xlab="x", ylab="y")
for(grp in unique(fitL$group)) {
ind <- which(fitL$group==grp)
ord <- order(fitL$x[ind,2])
lines(fitL$x[ind,2][ord],
fitL$y[ind][ord],
col=grey(0.9))
}
for(cls in 1:length(sfitL)) {
# The following implements the (3rd) method of
# Hanson & McMillan (2012) for simultaneous credible bands
# Generate coefficients from profile posterior
n <- 1e4
tau <- rgamma(n, shape=fitL$a[[cls]]/2, scale=2/fitL$b[[cls]])
muz <- matrix(rnorm(n*2, 0, 1),n,2)
mus <- (muz / sqrt(tau)) %*% chol(solve(fitL$s[[cls]]))
mu <- outer(rep(1,n), fitL$m[[cls]]) + mus

# Compute Mahalanobis distances
mhd <- rowSums(muz^2)

# Find the smallest 95% in terms of Mahalanobis distance
# I.e., a 95% credible region for mu
ord <- order(mhd, decreasing=TRUE)[-(1:floor(n*0.05))]
mu <- mu[ord,]
#Compute the 95% credible band
plotx <- seq(min(dat$x), max(dat$x), length.out=200)
ral <- apply(mu, 1, function(m) m[1] + m[2] * plotx)
rlo <- apply(ral, 1, min)
rhi <- apply(ral, 1, max)
```

```
rmd <- fitL$m[[cls]][1] + fitL$m[[cls]][2] * plotx

lines(plotx, rmd, col=cls, lty=2)
lines(plotx, rhi, col=cls)
lines(plotx, rlo, col=cls)
}
%dev.off()
```