MATH 640 Note Set I

Mark J. Meyer

January 10, 2018

GEORGETOWN UNIVERSITY

Georgetown College

Department of Mathematics and Statistics

Welcome to MATH 640!

Class Meetings

Thursdays from 6:30pm to 9:00pm, Walsh 495

Office Hours

Thursdays from 5:00pm to 6:00pm, St. Mary's 321 (me)

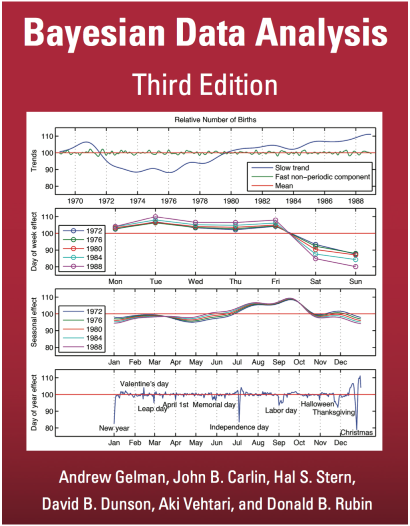
Drop-ins (if my door is open) and by appointment (me and Jim)

Contact

My e-mail: mjm556@georgetown.edu ($\sim$72 hour turn around)
Jim's e-mail: jwp60@georgetown.edu

Please pick up a copy of the syllabus!

Optional Text book



Bayesian Data Analysis

Third Edition

Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin

Suggested (it's a good reference)

Course Organization

Topics are organized into Three parts:

- Introduction to Bayesian Thought
  - Note Set I, ∼ BDA Chapter 1 and Appendix A
- Part I: Bayesian Theory and Direct Sampling
  - Note Set I, ∼ BDA Chapters 2—5, 14, 16
- Part II: Bayesian Analysis and Computation
  - Note Set II, ∼ BDA Chapters 6, 7, 10—12, 14—16

Introduction to Bayesian Thought

Chapter 1 and Appendix A

Introduction to Bayesian Thought Units

A. Introduction to Bayes

B. Probability and Inference

Unit A: Introduction to Bayes

Chapter 1

Introduction to Bayes

Consider an unknown parameter that we wish to estimate, call it θ

Introduction to Bayes

Consider an unknown parameter that we wish to estimate, call it θ

Question: How would you proceed to estimate and draw inference on this parameter?

Introduction to Bayes

Let $X_i, i = 1, \ldots, n$ be realizations from the population that is described by θ

Frequentist Perspective

We may estimate θ with $\bar{X} = 1/n \sum_i X_i$. Then, via the Central Limit Theorem, provided n is sufficiently large (and finite variance), we have that

$$\frac{\sqrt{n}\left(\bar{X} - \theta\right)}{\sigma} \xrightarrow{D} Z$$

where $Z \sim N(0, 1)$ and conduct inference accordingly

Introduction to Bayes

In practice, the CLT works by assuming we observe repeated samples from the population and that the distribution of the means of each successive sample approaches normality as the sample size approaches infinity

Thus we can describe the sampling distribution of $\bar{X}$ as

$$\bar{X} \mathrel{\dot\sim} N\left(\theta, \frac{\sigma^2}{n}\right)$$

From the sampling distribution, provided we have an estimate for σ^2, we can conduct inference about θ

Introduction to Bayes

- The sampling distribution is the basis for quantifying uncertainty in the frequentist paradigm
- All of the standard measures of uncertainty are derived from the sampling distribution
  - standard error:

    $$SE = \frac{s}{\sqrt{n}}$$

  - confidence interval:

    $$\bar{X} \pm z_{1-\alpha/2} SE$$

  - p-value: $2 * P\left(Z > \left|\frac{\bar{X}-\theta_0}{s}\right|\right)$
- Interpreting each of these measures of uncertainty requires consideration of the interpretation of the sampling distribution

Introduction to Bayes

From Gelman: an alternative is to consider what our experience/knowledge tells us and say that it forms a ‘belief structure’

Introduction to Bayes

From Gelman: an alternative is to consider what our experience/knowledge tells us and say that it forms a ‘belief structure’

As people gain experience/knowledge they (somehow) incorporate it into their belief structures

Introduction to Bayes

From Gelman: an alternative is to consider what our experience/knowledge tells us and say that it forms a ‘belief structure’

As people gain experience/knowledge they (somehow) incorporate it into their belief structures

Two key questions

Question: How do we formally quantify ‘belief’?
Question: How do we formally incorporate additional experience/knowledge?

Introduction to Bayes

From Gelman: an alternative is to consider what our experience/knowledge tells us and say that it forms a 'belief structure'

As people gain experience/knowledge they (somehow) incorporate it into their belief structures

Two key questions

Question: How do we formally quantify 'belief'?
Question: How do we formally incorporate additional experience/knowledge?

The Bayesian paradigm provides a coherent framework within which both of these questions are answered

Introduction to Bayes

According to Gelman, the basic steps of a Bayesian analysis are:

(1) quantify current beliefs via a *prior distribution*

(2) quantify information provided by data via the *likelihood*

(3) use Bayes' Theorem to update beliefs and form the *posterior distribution*

Introduction to Bayes

According to Gelman, the basic steps of a Bayesian analysis are:

(1) quantify current beliefs via a *prior distribution*

(2) quantify information provided by data via the *likelihood*

(3) use Bayes' Theorem to update beliefs and form the *posterior distribution*

These steps are somewhat backward from how a data analysis typically proceeds

Introduction to Bayes

There are several issues with the ordering of the first two steps in particular as

- it is hard to quantify a prior belief *before* knowing the form of the likelihood
- the likelihood will dictate the nature of the prior
- we may have an absence of current beliefs

Introduction to Bayes

Thus we may modify the steps as follows:

(1) identify a quantity of interest and collect data

(2) quantify information provided by data via the *likelihood*

(3) quantify current beliefs via a *prior distribution* (may be a lack of belief)

(4) use Bayes' Theorem to update beliefs and form the *posterior distribution*

As we'll see, the posterior distribution is the basis for all statements of statistical estimation and inference in Bayesian Statistics

Introduction to Bayes

Back to our unknown parameter, θ

Introduction to Bayes

Back to our unknown parameter, θ

Question: How do we quantify ‘belief’ about the parameter θ?

Introduction to Bayes

Back to our unknown parameter, θ

Question: How do we quantify 'belief' about the parameter θ?

- How do we assign relative weight or mass or probability to any given value?

Introduction to Bayes

Back to our unknown parameter, θ

Question: How do we quantify ‘belief’ about the parameter θ?

- How do we assign relative weight or mass or probability to any given value?

General idea is to treat θ as if it were a random variable

- Can then assign θ a distribution
- Use this distribution to assign relative weight or probability to different values of θ
- Just like we do with any random variable

Useful to interpret this distribution as a means to characterize uncertainty in our knowledge about θ

Introduction to Bayesian Thought

Note, we <u>do not</u> say that θ is a random variable

- the underlying parameter is, of course, constant

However, treating θ as a random variable provides a route to using distributions as a means to assign relative weight/probability

- Very convenient for a number of reasons:
  - distributions have nice properties
    - assign non-negative mass to potential values
    - integrate to 1
  - we know how to work with distributions
    - calculate moments, quantiles, etc.
  - there are loads of them!

Introduction to Bayesian Thought

Bayesian vs Frequentist in a nut-shell

Frequentists estimate the unknown parameter θ using the data X by first finding a statistic (a function of X that is usually also sufficient) and then build inferential procedures by describing the sampling distribution of the statistic

Bayesians estimate the unknown parameter θ using the data X by first estimating and/or sampling from the distribution of $\theta|X$ (potentially given some prior information) and then build statistics and inferential procedures using random draws from that distribution

It's ultimately a philosophical difference about estimation

Introduction to Bayesian Thought

Bayesian methods require a working knowledge of many concepts from probability

Thus before we proceed to developing Bayesian models, we require a thorough review of probability and inference

Unit B: Probability and Inference

Appendix A

Probability and Inference

As before, assume that θ is some unknown parameter

Marginal Distribution

The *marginal distribution* of θ is denoted $p(\theta)$ or $\pi(\theta)$

Further let X denote some other random variable

Joint Distribution

The *joint distribution* of θ and X is denoted as $p(\theta, X)$

Conditional Distribution

The *conditional distribution* of θ given X is denoted as $p(\theta|X)$

Probability and Inference

The conditional distribution is defined in terms of the joint and marginal distributions:

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)}$$

We can then also express the joint distribution in terms of the marginals and conditionals:

$$p(\theta, X) = p(X|\theta)p(\theta)$$

or

$$p(\theta, X) = p(\theta|X)p(X)$$

Probability and Inference

Notes on Conditional Distribution:

- Useful in probability for checking independence
- Has the effect of restricting the sample space
  - Useful for Bayesians, akin to restricting the parameter space of a model to what is informed by the data
- Valid for both discrete and continuous distributions
- $p(\theta|X)$ is a valid probability distribution, i.e.
  - $p(\theta|X) \geqslant 0$
  - $\sum_\theta p(\theta|X) = 1$ or $\int p(\theta|X)d\theta = 1$

Probability and Inference

Using the previous definitions, we get

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(X|\theta)p(\theta)}{p(X)}$$

Law of Total Probability

Let $A_1, A_2, \ldots$ be a partition of the sample space $\mathcal{S}$. Further let B be another element in $\mathcal{S}$. Then

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots = \sum P(B \cap A_i)$$

Noting that $P(B \cap A_i) = P(B|A_i)P(A_i)$, we have

$$P(B) = \sum P(B|A_i)P(A_i)$$

Probability and Inference

In terms of densities, the LTP can be expressed as

$$p(X) = \sum_{\theta} p(X|\theta)p(\theta) \text{ or } p(X) = \int p(X|\theta)p(\theta)d\theta$$

Combining with conditional probability expression from before gives us Bayes' Rule

Bayes' Rule

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\sum_{\theta} p(X|\theta)p(\theta)}$$

which forms the foundation of Bayesian estimation

Probability and Inference

Example I.1

An insurance company believes that people can be divided into two classes: those who are accident prone and those who are not. The company's statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability 0.4, whereas the probability decreases to 0.2 for a person who not accident prone. Assume that 30% of population is accident prone. Suppose a new policyholder has an accident within a year of purchasing a policy. What is probability that he or she is accident prone?

(from *A First Course in Probability*, 9^{th} *Ed.*, S. Ross)

Probability and Inference

Some additional concepts from probability:

Expectation

$$E(\theta) = \int \theta p(\theta) d\theta \text{ or } E(\theta) = \sum_{\theta} \theta p(\theta)$$

Variance

$$Var(\theta) = \int [\theta - E(\theta)]^2 p(\theta) d\theta \text{ or } Var(\theta) = \sum_{\theta} [\theta - E(\theta)]^2 p(\theta)$$

Further note, $Var(\theta) = E(\theta^2) - [E(\theta)]^2$

Probability and Inference

Often it is useful to express the mean and variance of a random variable θ in terms of the conditional mean given some related quantity X

Iterative expectation

$$E(\theta) = E[E(\theta|X)]$$

Proof: (see BDA or Casella-Berger)

Probability and Inference

We can also express the variance in terms of $\theta|X$

Iterative variance

$$Var(\theta) = E[Var(\theta|X)] + Var[E(\theta|X)]$$

Proof: (see BDA or Casella-Berger)

Probability and Inference

Example I.2

Suppose we wish to study the number of ponderosa pine trees in the Black Hills of South Dakota that are infected with the mountain pine beetle (a beetle that ultimately kills ponderosa pines). Let X be the number of infected ponderosa pines in a 10 square mile plot of forested land in the Black Hills. Further, let N be the number of ponderosa pines in that plot and p be the known probability that a randomly selected pine is infected. Finally, note that $X|N \sim Binom(N, p)$ and $N \sim Pois(\Lambda)$. Determine the following:

1. $E(X)$
2. $Var(X)$

Probability and Inference

Often, we may want to transform a probability distribution from one parameterization to another, thus taking the distribution from one parameter space to another parameter space

As multiple densities will be denoted, we let $p_\theta(\theta)$ and $p_\mu(\mu)$ denote the marginal distributions of θ and μ respectively

Further, let $\mu = g(\theta)$ be a one-to-one transformation

Probability and Inference

Must consider if the distribution is discrete or continuous

Discrete Transformation

$$p_{\mu}(\mu) = p_{\theta}\left[g^{-1}(\mu)\right]$$

Continuous Transformation

$$p_{\mu}(\mu) = p_{\theta}\left[g^{-1}(\mu)\right]|J|$$

where J is the Jacobian consisting of elements $\frac{\partial \mu_i}{\partial \theta_j}$ for the (i, j)th entry, if μ and θ are scalars, J is just $\frac{\partial}{\partial \theta} g^{-1}(\mu)$

Probability and Inference

Some common transformations

Log Transformation

Suppose $\theta \in (0, \infty)$. If $\mu = \log(\theta)$, then $\mu \in (-\infty, \infty)$.

Logit Transformation

Suppose $\theta \in (0, 1)$. If $\mu = \text{logit}(\theta)$, then $\mu \in (-\infty, \infty)$.
where $\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$

When $\theta \in (0, 1)$, the probit transformation can also be used,
$g(\theta) = \phi^{-1}(\theta)$

Probability and Inference

One useful transformation, particularly for randomly sampling from a distribution, is the Probability Integral Transform

Probability Integral Transform

Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then $Y \sim U(0, 1)$.

Corollary: Let $U \sim U(0, 1)$. Then, provided $F_X(x)$ is invertible, the random variable $W = F_X^{-1}(U)$ has distribution F_X.

Probability and Inference

Example I.3

Suppose $\mathcal{U} \sim \mathcal{U}(0,1)$ and let $W = -\frac{1}{\lambda}\log(1-\mathcal{U})$. What distribution does W have?

Probability and Inference

We will use a large number of distributions this semester

Many occur more or less only in Bayesian analysis—we will discuss these as they come up in the material

For now, we review some of the more common distributions we will encounter this semester

Probability and Inference

Uniform Distribution

A distribution where a variable is know to lie in an interval with equal probability across the interval, parameterized by $a, b \in \mathbb{R}$ with $a < b$

Notation: $\theta \sim U(a, b)$

$E(\theta) = \frac{a+b}{2}$

Density: $p(\theta) = \frac{1}{b-a}$

$Var(\theta) = \frac{(b-a)^2}{12}$

Support: $\theta \in [a, b]$

No mode

Useful R-functions: `runif()`, `dunif()`, `punif()`

Probability and Inference

Univariate Normal Distribution

A symmetric and unimodal distribution that is ubiquitous in statistics, parameterized by $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$

Notation: $\theta \sim N(\mu, \sigma^2)$

$E(\theta) = \mu$

Density: $p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\theta - \mu)^2\right]$

$Var(\theta) = \sigma^2$

Support: $\theta \in \mathbb{R}$

$Mode(\theta) = \mu$

Useful R-functions: `rnorm()`, `dnorm()`, `pnorm()`

Probability and Inference

Gamma Distribution

A general skewed distribution that is useful for incorporating current beliefs into many models, parameterized by $\alpha, \beta \in (0, \infty)$

Notation: $\theta \sim Gamma(\alpha, \beta)$

$E(\theta) = \frac{\alpha}{\beta}$

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

$Var(\theta) = \frac{\alpha}{\beta^2}$

Support: $\theta \in (0, \infty)$

$Mode(\theta) = \frac{\alpha-1}{\beta}, \alpha > 1$

Useful `R`-functions: `rgamma()`, `dgamma()`, `pgamma()`

Probability and Inference

Exponential Distribution

A special case of the Gamma, $Gamma(\alpha = 1, \beta)$, useful for the distribution of waiting times for the next event in a Poisson process, parameterized by $\beta \in (0, \infty)$

Notation: $\theta \sim Exp(\beta)$	$E(\theta) = \frac{1}{\beta}$
Density: $p(\theta) = \beta e^{-\beta\theta}$	$Var(\theta) = \frac{1}{\beta^2}$
Support: $\theta \in (0, \infty)$	$Mode(\theta) = 0$

Useful R-functions: `rexp()`, `dexp()`, `pexp()`

Probability and Inference

Beta Distribution

A distribution defined on the unit interval, it is generalized version of standard Uniform that is useful for incorporating current beliefs into models, parameterized by $\alpha, \beta \in (0, \infty)$

Notation: $\theta \sim Beta(\alpha, \beta)$

$E(\theta) = \frac{\alpha}{\alpha+\beta}$

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

$Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Support: $\theta \in (0, 1)$

$Mode(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$

Useful R-functions: `rbeta()`, `dbeta()`, `pbeta()`

Probability and Inference

Continuous Distributions in R

Example I.R1

In R, generate 1000 draws from a standard normal distribution and plot the empirical density. Then overlay the theoretical density of the standard normal and compare. Set the seed to 1212.

R tips:

- `set.seed()` sets the seed
- `rnorm()` generates random samples from a univariate normal
- `dnorm()` produces the pdf of a univariate normal
- `plot(density(...))` produces a smoothed empirical density based off of a random sample
- `curve()` can be used to plot functions

Probability and Inference

Poisson Distribution

A distribution that is commonly used to represent count data, parameterized by the rate $\lambda \in (0, \infty)$

Notation: $\theta \sim Pois(\lambda)$

$E(\theta) = \lambda$

Density: $p(\theta) = \frac{\lambda^\theta e^{-\lambda}}{\theta!}$

$Var(\theta) = \lambda$

Support: $\theta = 0, 1, 2, \ldots$

$Mode(\theta) = \lfloor \lambda \rfloor$

Useful R-functions: `rpois()`, `dpois()`, `ppois()`

Probability and Inference

Binomial Distribution

A distribution that is commonly used to represent the number of ‘successes’ in a sequence of n independent and identically distributed Bernoulli trials, parameterized by number of trials $n = 1, 2, 3, \ldots$ and probability of success $\pi \in [0, 1]$

Notation: $\theta \sim Bin(n, \pi)$ $E(\theta) = n\pi$

Density: $p(\theta) = \binom{n}{\theta}\pi^{\theta}(1-\pi)^{n-\theta}$ $Var(\theta) = n\pi(1-\pi)$

Support: $\theta = 0, 1, 2, \ldots, n$ $Mode(\theta) = \lfloor (n+1)\pi \rfloor$

Useful R-functions: `rbinom()`, `dbinom()`, `pbinom()`

Discrete Distributions in R

Example I.R2

In R, generate 1000 draws from a Poisson distribution with $\lambda = 3$ and plot the empirical density. Then, overlay the theoretical mass function and compare. Set the seed to 327.

R tips:

- `rpois()` generates random samples from a Poisson
- `dpois()` produces the pmf of a Poisson
- The argument `type = 'h'` along with the functions `plot()` and `points()` can be used to produce an appropriate mass function (use `x` of 0:15, `ylim` may need adjusting)
- Use `lines()` and `table()` to generate the empirical mass function

Probability and Inference

Distributions can be grouped into different families: scale, location, location-scale, etc.

An important family for Bayesians is the exponential family

Exponential Family

A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$p(x) = h(x)c(\theta) \exp \left[\sum_{i=1}^{k} w_i(\theta) t_i(x) \right]$$

where θ is potentially a vector of parameters and $t_i(x)$ are real-valued

Probability and Inference

Example I.4

Let $X \sim N(\mu, \sigma^2)$. Show that the family of pdfs where $\theta = (\mu, \sigma), -\infty < \mu < \infty, \sigma > 0$ is an exponential family.

Probability and Inference

Definition: Kernel
The *kernel* of a distribution is the part of a pdf or pmf that is a function of the random variable and the parameters. It cannot be simplified and is not part of the normalizing constant.

Kernel Recognition

If we can recognize distribution kernels, it can make a number of calculations easier in probability

For Bayesians, it will be useful for determining the distribution of $\theta|X$ as well as selecting reasonable prior beliefs

Probability and Inference

Example I.5

Let $\theta \sim \mathrm{Gamma}(\alpha, \beta)$. Using Kernel recognition, show that $E(\theta) = \alpha/\beta$.

Probability and Inference

Likelihood Function

Let $f(x_i|\theta)$ denote the pdf or pmf of the sample $\mathbf{X} = X_1, \ldots, X_n$. Assuming we observed $\mathbf{X}$, the likelihood is defined as

$$\mathcal{L}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where $\mathbf{x} = [x_1 \cdots x_n]$, the observed values of $\mathbf{X}$.

Note: in inference, the function is commonly written as $\mathcal{L}(\theta|\mathbf{x})$ to illustrate that θ is unknown and being estimated from the likelihood by the known, and presumably fixed, $\mathbf{x}$

Probability and Inference

Notes on the likelihood:

- Bayesian inference obeys what is called the *Likelihood Principle:*

  "For a given sample of data, any two probability models $p(x|\theta)$ that have the same likelihood function yield the same inference for θ"

- In modeling, we can rarely be confident that our chosen likelihood is correct, thus varying $\mathcal{L}(\mathbf{x}|\theta)$ is highly encouraged—and quite simple to do

Probability and Inference

Example I.6

Let $X_1, \ldots, X_n \sim \mathrm{Pois}(\lambda)$. Determine the likelihood function. If, instead, $X_1, \ldots, X_n \sim \mathrm{Pois}(\lambda_i)$, what would the likelihood look like?

Probability and Inference

Sufficient Statistics

If $p(\mathbf{X}|\theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a *sufficient statistic for* θ if, for every $\mathbf{x}$ in the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant as a function of θ.

In other words, the likelihood depends on $T(\mathbf{X})$ only through a function of both θ and $T(\mathbf{X})$.

Consider an exponential family:

$$p(x) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right]$$

Probability and Inference

Returning to Bayes' Rule

Bayes' Rule

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\sum_\theta p(X|\theta)p(\theta)} \text{ or } \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}$$

It is not necessary to evaluate $\sum_\theta p(X|\theta)p(\theta)$ or $\int p(X|\theta)p(\theta)d\theta$

Note that $p(X)$ does not depend on θ and if we assume X is fixed, it can be treated as a constant

Probability and Inference

Assuming $p(X)$ is a constant, means we can re-express the resulting density from Bayes' rule up to a constant of proportionality

Note: the operator $\propto$ means "proportional to"

Thus we get the foundation of Bayesian estimation

Basic Bayesian Model

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

Probability and Inference

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

Let's examine this component by component:

$p(\theta|X)$

The resulting distribution is called the *posterior distribution* (or to be more exact, the *unnormalized posterior density*). The ultimate goal of Bayesian Statistics is to determine and sample from the posterior. From there, we can develop statistics and conduct inference.

Probability and Inference

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

Let's examine this component by component:

$p(X|\theta)$

The first component of the RHS is the *Likelihood* function. This determined by what your data looks like (i.e. continuous or discrete) and how it was collected. Note that Gelman will use $p(X|\theta)$ to denote the likelihood but it is perfectly acceptable to use $\mathcal{L}(X|\theta)$ instead.

Probability and Inference

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

Let's examine this component by component:

$p(\theta)$

The second component of the RHS is the *prior distribution*. The prior represents our previous beliefs about the nature of θ. Note that it is only a function of θ. Selecting a prior requires many considerations, which we will explore later. Note that Gelman denotes the prior with $p(\theta)$ but $\pi(\theta)$ is also common.

Probability and Inference

Using the alternate notation discussed above, we may also express the basic Bayesian model as

$$p(\theta|X) \propto \mathcal{L}(X|\theta)\pi(\theta)$$

Some find this useful in distinguishing which components are which

Probability and Inference

Given a model, we may be interested in prediction based off a new observation

Suppose we observe a new data point, $\tilde{x}$

Posterior predictive distribution

Conditional on the data we already observed,

$$p(\tilde{x}|X) = \int p(\tilde{x}, \theta|X) d\theta = \int p(\tilde{x}|\theta, X) p(\theta|X) d\theta$$

which, assuming $X \perp \tilde{x}$ given θ, gives

$$p(\tilde{x}|X) = \int p(\tilde{x}|\theta) p(\theta|X) d\theta$$

Probability and Inference

The basic model can easily be expanded to account for additional parameters

Recall the multiplication rule

Multiplication Rule

Let $A_1, A_2, A_3, \ldots, A_n$ be an arbitrary number of events in a sample space. The probability of the intersection of the events can be expressed as

$$\begin{aligned} P(A_1, A_2, A_3, \ldots, A_n) &= P(A_1|A_2, A_3, \ldots, A_n) P(A_2|A_3, \ldots, A_n) \\ &\quad \times P(A_3|A_4, \ldots, A_n) \cdots P(A_{n-1}|A_n) P(A_n) \end{aligned}$$

Probability and Inference

The multiplication rule applies to densities as well

Let θ be an unknown parameter of interest and let γ be a secondary, unknown parameter

Further, let X be data collected to study θ, the basic model is then

$$p(\theta|X) \propto p(X|\theta,\gamma)p(\theta|\gamma)p(\gamma)$$

or

$$p(\theta|X) \propto \mathcal{L}(X|\theta,\gamma)\pi(\theta|\gamma)\pi(\gamma)$$

Probability and Inference

$$p(\theta|X) \propto \mathcal{L}(X|\theta, \gamma)\pi(\theta|\gamma)\pi(\gamma)$$

Note that we now have two priors: $\pi(\theta|\gamma)$ and $\pi(\gamma)$

Further note that $\pi(\theta|\gamma)$ is a function of the second unknown parameter

When this is the case, γ is called a hyper-parameter and $\pi(\gamma)$ a hyper-prior (i.e. a parameter for the prior distribution and a prior on that parameter)

Part I: Fundamentals of Bayesian Inference

BDA Chapters 2-5, 14, and 16

Part I Units

A. Single-parameter Models

B. Multi-parameter Models

C. Normal Approximation

D. Hierarchical Models

Unit A: Single-parameter Models

Chapter 2

I.A Single-parameter Models

Returning to the basic Bayesian model, let's consider a simple example

Recall the form of the basic model:

$$p(\theta|X) \propto \mathcal{L}(X|\theta)\pi(\theta)$$

Example I.7

Let's build a basic Bayesian model to study the number of ponderosa pine trees in Black Hills of South Dakota that are infected with the mountain pine beetle.

Question: What components do we need?

I.A Single-parameter Models

Let X be the number of ponderosa pines infected out of n total pines in the region of study. Further, let θ be the true proportion of infected pines.

We could assume that $X|\theta \sim Bin(n, \theta)$

Thus the likelihood is

$$\mathcal{L}(X|\theta) = \binom{n}{X} \theta^{X} (1 - \theta)^{n-X}$$

I.A Single-parameter Models

Note that n is an attribute of our data and thus can be considered fixed

Thus the leading term is a normalizing constant and the likelihood can be expressed as

$$\mathcal{L}(X|\theta) \propto \theta^{X}(1-\theta)^{n-X}$$

Question: What shall we select for the prior on θ?

I.A Single-parameter Models

While we have many choices, let's take a simple one for now:

$$\theta \sim \mathcal{U}(0, 1)$$

Thus,

$$\pi(\theta) = 1$$

Multiplying by the likelihood gives the posterior

$$p(\theta|X) \propto \theta^{X}(1-\theta)^{n-X} \cdot 1 = \theta^{X}(1-\theta)^{n-X}$$

Question: What distribution does $\theta|X$ have?

I.A Single-parameter Models

Try adding and subtracting one to each exponent:

$$p(\theta|X) \propto \theta^{X+1-1}(1-\theta)^{n-X+1-1}$$

Question: What distribution has this as its kernel?

I.A Single-parameter Models

Try adding and subtracting one to each exponent:

$$p(\theta|X) \propto \theta^{X+1-1}(1-\theta)^{n-X+1-1}$$

Question: What distribution has this as its kernel?

It is the kernel of a $Beta(X+1, n-X+1)$ pdf

I.A Single-parameter Models

Try adding and subtracting one to each exponent:

$$p(\theta|X) \propto \theta^{X+1-1}(1-\theta)^{n-X+1-1}$$

Question: What distribution has this as its kernel?

It is the kernel of a $Beta(X+1, n-X+1)$ pdf

Thus, $\theta|X \sim Beta(X+1, n-X+1)$

I.A Single-parameter Models

With a closed form posterior and given some data, we can do many things without taking random samples:

1. Plot the theoretical density
2. Directly estimate posterior summaries:
   - Center: mean, median
   - Quantiles: 25%-ile, 75%-ile
   - Spread: variance, standard deviation, IQR
3. Calculate the posterior predictive distribution

I.A Single-parameter Models

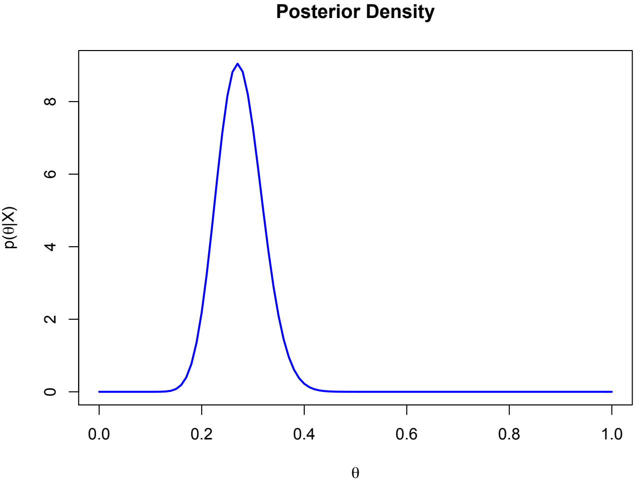
Example I.8

Suppose we sample 100 pine trees and determine that 27 have been infected with the mountain pine beetle.

Using this data, let's plot the posterior density, find posterior summaries, and determine the posterior predictive distribution

I.A Single-parameter Models

First consider plotting the theoretical density:



See Part I Code File on Canvas for R code

I.A Single-parameter Models

Next, recall the mean and variance for the Beta distribution:

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \text{ and } Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Also note the posterior mode:

$$Mode(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$$

Noting that $\alpha = X + 1$ and $\beta = n - X + 1$, what are the mean, variance, and model in the context of the model? What are they in the context of the data?

I.A Single-parameter Models

Now let's suppose a new tree is sampled. Based on our model, what is the probability it is infected with the mountain pine beetle?

$$\Pr(\tilde{x} = 1|X) = \int_0^1 \Pr(\tilde{x} = 1|\theta, X)p(\theta|X)d\theta$$

Note that $\tilde{X}|\theta \sim Bin(n, \theta)$, thus $\Pr(\tilde{x} = 1|\theta) = \theta$ and

$$\Pr(\tilde{x} = 1|X) = \int_0^1 \theta p(\theta|X)d\theta = E(\theta|X) = \frac{X+1}{n+2}$$

This should make sense if each trial is truly independent

I.A Single-parameter Models

Looking back, note the complexity of the variance calculation

Further note the standard deviation in context is

$$sd(\theta|X) = \sqrt{\frac{(X+1)(n-X+1)}{(n+2)^2(n+3)}}$$

which with data is ≈ 0.044

Question: what does this tell us about the spread?

I.A Single-parameter Models

Question: What percentage of the distribution falls within one standard deviation of the mean? How about two standard deviations?

I.A Single-parameter Models

Question: What percentage of the distribution falls within one standard deviation of the mean? How about two standard deviations?

First attempt: Empirical rule

I.A Single-parameter Models

Question: What percentage of the distribution falls within one standard deviation of the mean? How about two standard deviations?

First attempt: Empirical rule

Empirical Rule

For a symmetric and unimodal distribution, roughly 68% of the distribution falls within one standard deviation of the mean while 95% falls within two standard deviations (and 99.7% falls within three)

I.A Single-parameter Models

Only works for symmetric and unimodal distributions

Empirical Rule is only an approximation

I.A Single-parameter Models

Only works for symmetric and unimodal distributions

Empirical Rule is only an approximation

Second attempt: Chebyshev's Inequality

I.A Single-parameter Models

Only works for symmetric and unimodal distributions

Empirical Rule is only an approximation

Second attempt: Chebyshev's Inequality

Chebychev's Inequality

For distributions that are not symmetric and unimodal, at least 75% of the distribution falls within 2 standard deviations of the mean while 88.9% of the distribution falls within 3 standard deviations of the mean.

Ultimately, neither is satisfactory for a general approach to interval construction in the Bayesian context

An alternative approach is to rely on random samples and build summaries using those samples—particularly useful for intervals

Credible Interval

An interval corresponding to the middle $100(1-\alpha)\%$ of a distribution taken from posterior samples (Gelman calls this the *posterior interval*)

I.A Single-parameter Models

Example I.R3

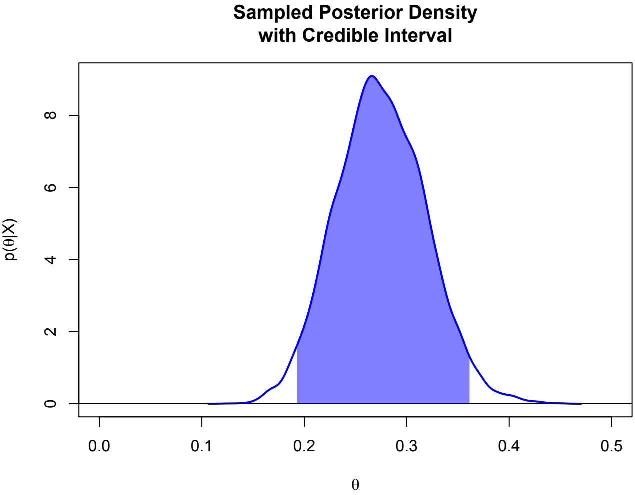
Generate 10000 from the posterior distribution and find the 95% credible interval. Recall that $X = 27$ and $n = 100$. Also plot the posterior distribution based off the samples. Set the seed to 526.

`R` tips:

- Use `plot(density(samples), ...)` to plot the density
- The function `quantile()` is useful for obtaining the bounds of the credible interval, the argument `probs = ...` can be used to specify desired quantiles

I.A Single-parameter Models

We can also visualize the credible interval:



See Part I Code File on Canvas for `R` code

I.A Single-parameter Models

Backing up, how did we decide the prior on θ? Are there other priors we could select?

A key question to ask: how does the choice of the prior affect the posterior?

Let's formalize the prior selection process

I.A Single-parameter Models

First consider two types of prior distributions: informative and non-informative

Informative Prior

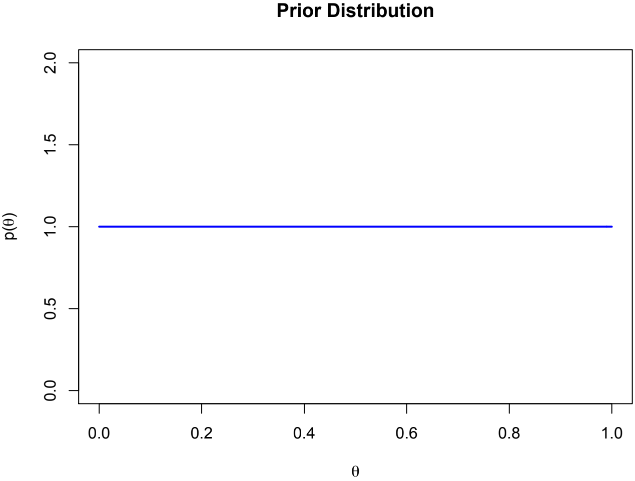
Informative priors provide information about the location of the parameter of interest, i.e. the mass of the prior is *centralized*, that information can be a lot or a little depending on its shape

Non-Informative Prior

Non-informative priors give no information about the location of the parameter of interest, i.e. the mass of the prior is *de-centralized*

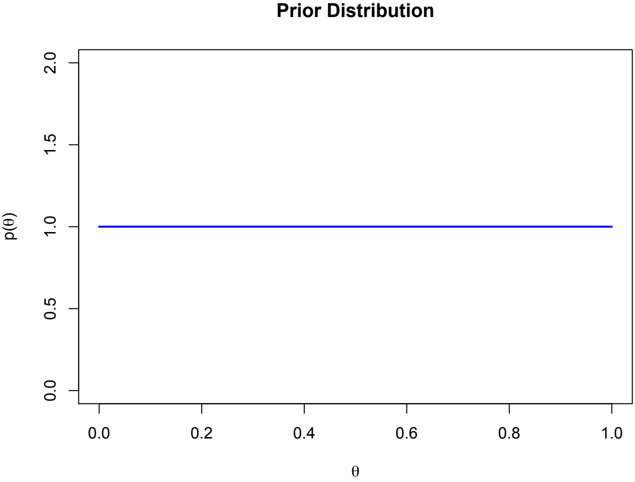
I.A Single-parameter Models

Consider the prior from the example: $\theta \sim \mathcal{U}(0,1)$



I.A Single-parameter Models

Consider the prior from the example: $\theta \sim \mathcal{U}(0, 1)$



Question: Is this prior informative or non-informative?

I.A Single-parameter Models

Not only is this prior non-informative (within the sample space), it is also *flat*

Flat Prior

A flat prior is a prior that when graphed is flat or nearly flat, typically non-informative

Not all non-informative priors are flat (as we will see later)

I.A Single-parameter Models

Question: Given a likelihood, how do we select a prior (informative or not)?

I.A Single-parameter Models

Question: Given a likelihood, how do we select a prior (informative or not)?

The approach from before: pick a prior that gives us a well behaved, recognizable posterior

We can use this again to pick a different kind of prior

I.A Single-parameter Models

Consider, again, the likelihood:

$$\mathcal{L}(X|\theta) \propto \theta^{X}(1-\theta)^{n-X}$$

I.A Single-parameter Models

Consider, again, the likelihood:

$$\mathcal{L}(X|\theta) \propto \theta^{X}(1-\theta)^{n-X}$$

This kernel should remind us of the Beta distribution

I.A Single-parameter Models

Consider, again, the likelihood:

$$\mathcal{L}(X|\theta) \propto \theta^{X}(1-\theta)^{n-X}$$

This kernel should remind us of the Beta distribution

Thus we can select our prior to have the form

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

I.A Single-parameter Models

Our posterior distribution then has the form

$$p(\theta|X) \propto \theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}$$

I.A Single-parameter Models

Our posterior distribution then has the form

$$p(\theta|X) \propto \theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}$$

Question: What distribution is this?

I.A Single-parameter Models

Our posterior distribution then has the form

$$p(\theta|X) \propto \theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}$$

Question: What distribution is this?

Our choice of α, β dictates how informative (or not) the prior is

I.A Single-parameter Models

Our posterior distribution then has the form

$$p(\theta|X) \propto \theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}$$

Question: What distribution is this?

Our choice of α, β dictates how informative (or not) the prior is

If we have prior knowledge, we can incorporate that here

I.A Single-parameter Models

Also note that posterior distribution follows the same parametric form as the prior, i.e. they're both Beta distributions—a property known as *conjugacy*

We say that the beta prior distribution is a *conjugate family* for the binomial likelihood

This binomial "+" beta "=" beta relationship is just one instance of a special, mathematically convenient class of priors that are called *conjugate families*

They are convenient because the posterior follows a known parametric form

I.A Single-parameter Models

Conjugate Priors

If $\mathcal{F}$ is a class of sampling distributions $\mathcal{L}(y|\theta)$ and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta|y) \in \mathcal{P} \text{ for all } \mathcal{L}(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

We are most interested in *natural* conjugate prior families

Natural Conjugate Priors

Natural conjugate priors arise by taking $\mathcal{P}$ to be the set of all densities having the same functional form as the likelihood

I.A Single-parameter Models

Probability distributions belonging to an *exponential family* have a *natural* conjugate prior distribution

Suppose we take a distribution with one parameter:

$$p(x_i|\theta) = h(x_i)c(\theta)\exp\left[w(\theta)t(x_i)\right]$$

The likelihood is then

$$\begin{aligned}\mathcal{L}(X|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} h(x_i)c(\theta)\exp\left[w(\theta)t(x_i)\right] \\ &= \left[\prod_{i=1}^{n} h(x_i)\right] c(\theta)^n \exp\left[w(\theta)\sum_{i=1}^{n} t(x_i)\right]\end{aligned}$$

I.A Single-parameter Models

Thus the likelihood can be expressed as

$$\mathcal{L}(X|\theta) \propto c(\theta)^n \exp\left[w(\theta)T(X)\right]$$

where $T(X) = \sum_{i=1}^{n} t(x_i)$ is a sufficient statistic for θ

If we specify the prior to be

$$p(\theta) \propto c(\theta)^\eta \exp\left[w(\theta)\nu\right]$$

The posterior is then

$$p(\theta|X) \propto c(\theta)^{\eta+n} \exp\{w(\theta)[\nu + T(X)]\}$$

Thus the choice of prior is conjugate

I.A Single-parameter Models

Some additional notes on conjugate priors

- Conjugate priors are computationally convenient
- Can be interpreted as additional data
- They are a good starting point for modeling
- Conjugate priors may not exist for more complicated models

However, nonconjugate priors do not pose any new conceptual problems and can often provide more flexible choices in modeling

The trade-off comes in the form of more costly computation, but nonconjugate priors can lead to better models

I.A Single-parameter Models

Example I.9

Suppose we observe y, a single data point, from a normal distribution parameterized by a mean θ and variance σ^2. Further assume that (however unlikely) that σ^2 is known and therefore fixed. Let's build a basic model for θ using a conjugate prior.

Question: What if the mean is known but now the variance is unknown?

I.A Single-parameter Models

Inverse-Gamma

If $\theta^{-1} \sim \mathrm{Gamma}(\alpha, \beta)$, then θ is Inverse-Gamma which can be used as the conjugate distribution for the normal variance.

Notation: $\theta \sim \mathrm{IG}(\alpha, \beta)$

$E(\theta) = \frac{\beta}{\alpha-1}$

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

$Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$

Support: $\theta \in (0, \infty)$

$Mode(\theta) = \frac{\beta}{\alpha+1}$

Useful `R`-functions: `rinvgamma()` and `dinvgamma()` in the package `MCMCpack`

I.A Single-parameter Models

Inverse-Chi-Square

A special-case of the Inverse-Gamma with $\alpha = \nu/2$ and $\beta = 1/2$.

Notation: $\theta \sim Inv\text{-}\chi^2_\nu$

$E(\theta) = \frac{1}{\nu-2}$

Density: $p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{-(\frac{\nu}{2}+1)} e^{-\frac{1}{2\theta}}$

$Var(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}$

Support: $\theta \in (0, \infty)$

$Mode(\theta) = \frac{1}{\nu+2}$

Useful R-functions: `rinvchisq()` and `dinvchisq()` in the package `geoR` or Inverse-Gamma functions with appropriate parameters

I.A Single-parameter Models

Scaled Inverse-Chi-Square

If $X \sim \chi^2_\nu$ then $\theta = \nu s^2/X$ is a scaled inverse-chi-square distributed random variable. Or an inverse-gamma with $\alpha = \nu/2$ and $\beta = \frac{\nu}{2}s^2$

Notation: $\theta \sim \mathrm{Inv}\text{-}\chi^2(\nu, s^2)$

$E(\theta) = \frac{\nu}{\nu-2}s^2$

Density: $p(\theta) = \frac{\frac{\nu}{2}^{\nu/2}}{\Gamma(\nu/2)} s^\nu \theta^{-(\frac{\nu}{2}+1)} e^{-\frac{\nu s^2}{2\theta}}$

$Var(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)}s^4$

Support: $\theta \in (0, \infty)$

$Mode(\theta) = \frac{\nu}{\nu+2}s^2$

Useful R-functions: `rinvchisq()` and `dinvchisq()` in the package `geoR` with `scale` argument or Inverse-Gamma functions with appropriately defined parameters

I.A Single-parameter Models

Example I.10

Suppose instead we observe y_i, $i = 1, \ldots, n$, from a normal distribution parameterized by a known mean θ and unknown variance σ^2. In this case, we can consider θ. Once again, let's build a basic model for θ using a conjugate prior.

I.A Single-parameter Models

Conjugate and informative priors:

- Conjugacy can be used to select informative priors
  - All depends on the choice of the hyper-parameters (i.e. the parameters of the prior distribution)
- One way is to use a previous study's results as your prior
  - As the saying goes, "yesterday's posterior is today's prior"

Example I.11

Suppose we conduct a second study in the Black Hills to determine the proportion of ponderosa pines infected with the mountain pine beetle. Our prior could now be $\pi(\theta) \sim Beta(X + 1, n - X + 1)$ or $\pi(\theta) \sim Beta(28, 74)$ since $X = 27$ and $n = 100$.

I.A Single-parameter Models

Noninformative Priors

If no previous study has been conducted, it is hard to construct an informative prior. Further, there is a desire to have a prior that is guaranteed to play a minimal role in the posterior distribution.

Thus the need for vague, flat, or noninformative priors

The idea here is to “let the data speak for itself” so that inference is unaffected by information external to the current data

I.A Single-parameter Models

Thus far we have only considered prior distributions that were also valid pdfs

Proper Priors

A prior density, $\pi(\theta)$, is called *proper* if it does not depend on the data and integrates (or sums) to 1

But we don't necessarily need the prior to be a valid pdf/pmf

Improper Priors

A prior, $\pi(\theta)$, that depends on the data or does not integrate/sum to 1 or both (though typically not the former)

I.A Single-parameter Models

An improper prior may lead to a proper posterior

Using an improper prior

When using an improper prior, we may still have a proper posterior. If that is the case, then there is no harm in using the improper prior. Care must be taken, however, to ensure the posterior is indeed proper when implementing an improper prior.

I.A Single-parameter Models

Example I.12

Consider again the example where y_i, $i = 1, \dots, n$ is Normal with known mean θ and unknown mean σ^2. Suppose we take the prior $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$.

Questions:

1. Is the prior proper or improper?
2. Will it lead to a proper posterior?

I.A Single-parameter Models

Question: How do we define noninformative priors?

- Can use a flat prior
  - Like we did in the example with the ponderosa pines
  - Priors that are constants, for example
- Can select a diffuse prior
  - Similar to flat, but may still have regions with slightly higher mass
  - Define these by using conjugate priors and picking hyper-parameters to ensure a relatively flat density
- Can use Jeffreys' prior

I.A Single-parameter Models

Jeffreys' Invariance Principle

Consider a one-to-one transformation of the parameter θ: $\phi = h(\theta)$. By the transformation of variables, the prior density $\pi(\theta)$ is equivalent, in terms of expressing beliefs, to the prior density on ϕ:

$$\pi(\phi) = \pi(\theta) \left| \frac{d\theta}{d\phi} \right| = \pi(\theta) |h'(\theta)|^{-1}$$

I.A Single-parameter Models

Any rule for determining the prior density $\pi(\theta)$ should yield an equivalent result if applied to the transformed parameter

In other words, regardless of how we select a prior, if it is invariant, then the resulting models should be equivalent

I.A Single-parameter Models

This principle leads us to define a noninformative prior density

Jeffreys' Prior

A noninformative prior for θ is given by

$$\pi(\theta) \propto [J(\theta)]^{1/2}$$

where $J(\theta)$ is the *Fisher information* for θ defined as

$$J(\theta) = E\left[\left(\frac{d\ell(y|\theta)}{d\theta} \right)^2 \middle| \theta \right] = -E\left[\frac{d^2\ell(y|\theta)}{d\theta^2} \middle| \theta \right]$$

where $\ell(y|\theta) = \log\left[\mathcal{L}(y|\theta)\right]$

I.A Single-parameter Models

Jeffreys' prior follows Jeffreys' principle

To see this, evaluate $J(\phi)$ at $\theta = h^{-1}(\phi)$:

$$\begin{aligned} J(\phi) &= -E\left[\frac{d^2\ell(y|\phi)}{d\phi^2}\right] \\ &= -E\left[\frac{d^2\ell\{y|\theta = h^{-1}(\phi)\}}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right] = J(\theta)\left|\frac{d\theta}{d\phi}\right|^2 \end{aligned}$$

Thus, $J(\phi)^{1/2} = J(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|$

Example I.13

Consider again our investigation of the infection of ponderosa pines in the Black Hills of South Dakota with the mountain pine beetle. As before, we assume $X|\theta \sim Bin(n, \theta)$. Determine a noninformative prior for θ using Jeffreys' prior.

Note: we already have a noninformative and flat prior for θ, namely $\theta \sim U(0, 1)$.

Question: Will Jeffreys' prior result in the same noninformative prior or a different one?

I.A Single-parameter Models

Example I.14

Now find the posterior using Jeffreys' prior for θ given that $X|\theta \sim Bin(n, \theta)$. Compare the theoretical results of this posterior to what we found in Example I.8.

Example I.R4

Next, assume as before that we observed $n = 100$ ponderosa pine trees with $X = 27$ infected trees. Compare the posterior densities, posterior means, and 95% credible intervals that result from the two different models. Use the same seed as before, 526.

I.A Single-parameter Models

Difficulties with noninformative priors:

1. Searching for a prior that is always vague may be misguided

2. For many problems, there is no clear chose of a vague prior

3. Further difficulties can arise

I.A Single-parameter Models

An alternative to noninformative priors is to use a weakly informative prior

Weakly Informative Prior

A prior distribution is *weakly informative* if it is proper but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available.

Gelman suggests this approach as opposed to “trying to model complete ignorance”

I.A Single-parameter Models

Weakly informative priors may make more sense in certain models, depending on parameterization

Consider the following:

Example I.15

Suppose we wish to fit a regression model: $y_i = \alpha + \beta x_i$. We know that β can take on a range of values, but we might think a-priori that it is centered at 0. A weakly informative prior on β could then be centered at 0 but with a large variance, thus giving a diffuse prior that is still centered around 0.

Note: the variance needs to be large relative the scale of x_i

I.A Single-parameter Models

Constructing Weakly Informative Priors

Below are two principles, from different directions, for constructing weakly informative priors

- Start with a noninformative prior and add enough information so inferences are constrained to be reasonable
- Start with a highly informative prior and broaden it to account for uncertainty in prior beliefs and in historically based priors

I.A Single-parameter Models

Ultimately, prior selection can be a bit of art form and depends on the model one is trying to implement

But all priors should follow the *symmetry principle*

Symmetry Principle

The prior distribution should not pull inferences in any predetermined direction

We should also always check the sensitivity of our models to prior specification—known as a sensitivity analysis

Unit B: Multi-parameter Models

BDA Chapters 3 and 14

I.B Multi-parameter Models

Suppose we've collected data X whose likelihood depends on the parameters θ and γ

The Bayesian model can easily extend to multiple parameters

Let's consider the joint posterior of both these variables:

$$p(\theta, \gamma | X) \propto \mathcal{L}(X|\theta, \gamma)\pi(\theta, \gamma)$$

Here we have a joint prior for θ and γ

I.B Multi-parameter Models

$$p(\theta, \gamma | X) \propto \mathcal{L}(X | \theta, \gamma) \pi(\theta, \gamma)$$

Many statistical problems involve multiple parameters

Typically, inference is drawn on a model one parameter at time

Thus a goal of a Bayesian analysis is to obtain the marginal posterior distribution for each parameter which can obtain by integrating out the unknowns that are not of immediate interest

I.B Multi-parameter Models

In many problems, we do not conduct inference on every parameter

Nuisance Parameter

Parameters that are of no inferential interest but must still be estimated as part of the model are often called *nuisance parameters*

Example I.16

Suppose $y_i \sim N(\mu, \sigma^2)$. Often when working with the normal model we are primarily interested in conducting inference on the μ and rarely on σ^2—yet we still need to estimate it. Thus σ^2 is a nuisance parameter.

I.B Multi-parameter Models

Consider again our hypothetical two parameter model

$$p(\theta, \gamma | X) \propto \mathcal{L}(X | \theta, \gamma)\pi(\theta, \gamma)$$

We may only be interested in inference on θ

Averaging over Nuisance Parameters

We could obtain the marginal for θ by averaging over γ:

$$p(\theta | X) = \int p(\theta, \gamma | X) d\gamma$$

I.B Multi-parameter Models

We can express the joint posterior as the product of conditional posterior distributions given γ, thus

$$p(\theta|X) = \int p(\theta|\gamma, X)p(\gamma|X)d\gamma$$

Thus the marginal posterior of θ given X is a mixture of the conditional posterior distributions for $\theta|\gamma, X$ weighted by the marginal posterior for γ

Note: we rarely explicitly evaluate this integral

I.B Multi-parameter Models

But this representation suggests an important practical strategy for constructing and computing multi-parameter models

Marginal-Conditional Approach

1. First, draw γ from its marginal posterior, $p(\gamma|X)$
2. Second, draw θ from its conditional posterior distribution, $p(\theta|\gamma, X)$, given the drawn value of γ

I.B Multi-parameter Models

Example I.17

Suppose y_i, $i = 1, \ldots, n$ are iid $N(\mu, \sigma^2)$ where both parameters are unknown. The model we wish to estimate is then

$$p(\mu, \sigma^2|y) \propto \mathcal{L}(y|\mu, \sigma^2)\pi(\mu, \sigma^2)$$

Suppose we select the noninformative prior

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

What does our joint posterior look like?

I.B Multi-parameter Models

Recall that to use the approach of averaging over nuisance parameters, we need the posterior to factor into the form:

$$p(\theta|\gamma, X)p(\gamma|X)$$

In the context of this example, we're looking to factor the posterior like so:

$$p(\mu|\sigma^2, y)p(\sigma^2|y)$$

How do we accomplish this?

I.B Multi-parameter Models

The factorized form of the posterior is then

$$p(\mu, \sigma^2|y) \propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2\right] \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right]$$

Questions:

1 Which term or terms represent the conditional posterior of μ given σ^2 and y?

2 Which term or terms are need to find the marginal posterior of σ^2 given y?

I.B Multi-parameter Models

The factorized form of the posterior is then

$$p(\mu, \sigma^2|y) \propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2\right] \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right]$$

Questions:

1 Which term or terms represent the conditional posterior of μ given σ^2 and y?

2 Which term or terms are need to find the marginal posterior of σ^2 given y?

Let's handle these one at time

I.B Multi-parameter Models

1 Which term or terms represent the conditional posterior of μ given σ^2 and y?

The conditional posterior distribution is

$$\begin{aligned} p(\mu|\sigma^2, y) &\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] \\ &= \exp\left[-\frac{1}{2(\sigma^2/n)}(\mu-\bar{y})^2\right] \end{aligned}$$

which is the kernel of a normal distribution with mean $\bar{y}$ and variance σ^2/n

Thus, $\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n)$

I.B Multi-parameter Models

2 Which term or terms are need to find the marginal posterior of σ^2 given y?

All of them

To find the marginal posterior distribution, we must average the joint distribution over μ:

$$p(\sigma^2|y) \propto \int (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2\right] \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] d\mu$$

Note that, with respect to μ, the first two terms are constants, thus we only need to determine the integral of the last term

I.B Multi-parameter Models

Thus we have factored the joint posterior density as the product of conditional and marginal posterior posterior densities

This gives us a strategy for drawing samples from the joint posterior distribution:

1. Draw $\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2)$
2. Draw $\mu|\sigma, y \sim N(\bar{y}, \sigma^2/n)$

Note: For step 2, we use the simulated value of σ^2, one simulated σ^2 for each simulated μ

We call this approach the *marginal-conditional* sampling method

Alternatively, we can derive the marginal posterior of $\mu|y$

Standard t-distribution

The standard t-distribution arises by taking $Z \sim N(0,1)$ and $Q \sim \chi^2_n$. Then, $T = Z/(\sqrt{Q}/n) \sim t_n$. It was discovered by a Guinness brew master who wanted a distribution to better conduct quality control.

Notation: $\theta \sim t_n$

Density: $p(\theta) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{\theta^2}{n}\right)^{-(n+1)/2}$

Support: $\theta \in \mathbb{R}$

$E(\theta) = 0$

for $n > 1$

$Var(\theta) = \frac{n}{n-2}$

for $n > 2$

$Mode(\theta) = 0$

Useful R-functions: `rt()`, `dt()`, `pt()`

I.B Multi-parameter Models

General t-distribution

The general t-distribution arises by performing a location and scale transformation to the a standard t random variable.

Notation: $\theta \sim t_n(\mu, \sigma^2)$

Density:

$$p(\theta) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi\sigma^2}} \left[1 + \frac{1}{n}\frac{(\theta-\mu)^2}{\sigma^2}\right]^{-(n+1)/2}$$

Support: $\theta \in \mathbb{R}$

$E(\theta) = \mu$
for $n > 1$

$Var(\theta) = \frac{n}{n-2}\sigma^2$
for $n > 2$

$Mode(\theta) = \mu$

Useful R-functions: sample from `rt()` and then scale by σ and shift by μ

I.B Multi-parameter Models

As we are typically interested in inference on μ, we must find the posterior marginal of μ given y which can be found by integrating σ^2 out of the joint posterior:

$$\begin{aligned} p(\mu|y) &= \int_0^\infty p(\mu, \sigma^2|y)\,d\sigma^2 \\ &\propto \int_0^\infty (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\bar{y}-\mu)^2\right\}\right] d\sigma^2 \end{aligned}$$

First, recall the form of the Inverse-Gamma density:

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

I.B Multi-parameter Models

With some work, we've shown that

$$\begin{aligned} p(\mu|y) &\propto \left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]^{-n/2} \\ &\propto \left[1 + \frac{1}{n-1}\frac{(\mu-\bar{y})^2}{s^2/n}\right]^{-(n-1+1)/2} \end{aligned}$$

which is the kernel of the a t distribution with $n-1$ degrees of freedom, location μ, and scale σ^2

Thus, the marginal posterior distribution of μ is then

$$\mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

I.B Multi-parameter Models

Combined with the marginal posterior for σ^2, we can now independently sample from both marginals:

1. Draw samples from $\sigma^2 | y \sim \text{Inv-}\chi^2(n-1, s^2)$
2. Draw samples from $\mu | y \sim t_{n-1}(\bar{y}, s^2/n)$

Unlike with the previous approach, the order doesn't matter here

We can then construct summaries based on posterior samples as in the single-parameter case

I.B Multi-parameter Models

Further, by using the noninformative joint prior, $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$, we've shown that the posterior distribution of μ has the form

$$\left.\frac{\mu - \bar{y}}{s/\sqrt{n}}\right| y \sim t_{n-1}$$

the standard t-distribution with $n-1$ degrees of freedom

Note the similarity to the sampling distribution result:

$$\left.\frac{\bar{y} - \mu}{s/\sqrt{n}}\right| \mu \sim t_{n-1}$$

I.B Multi-parameter Models

The posterior predictive distribution, $p(\tilde{y}|y)$, can be found in an analogous fashion to the marginal posterior for μ

Direct calculation requires performing the following integration:

$$p(\tilde{y}|y) = \int\int p(\tilde{y}|\mu, \sigma^2, y)p(\mu, \sigma^2|y)d\mu d\sigma^2$$

From this, it can be shown that

$$\tilde{y}|y \sim t_{n-1}\left(\bar{y}, [1 + 1/n]^{1/2}s\right)$$

Can also simulate from $\tilde{y} \sim N(\mu, \sigma^2)$ given poster draws for μ, σ^2

I.B Multi-parameter Models

Example I.R5

Consider data from the HERS study, randomized trial of the effects of hormone replacement therapy on coronary heart disease. We are interested in modeling the change in total cholesterol between the end of the study and baseline. Differences are often normally distributed, so determine the marginal posterior distributions of the resulting model parameters assuming that change in total cholesterol is normal and using the noninformative prior. Generate 10000 draws from the posterior density and produce relevant summaries. Set the seed to 3.

`R` Tips:

- Data can be found in the file `hers.txt` under Datasets
- Useful functions: `rt()`, `rinvgamma()` from package `MCMCpack` with `shape` and `scale` defined accordingly

I.B Multi-parameter Models

Example I.R6

Continuing with the HERS data, now compare the results from Example I.R5 to those obtained by using the marginal-conditional strategy. Recall that the marginal and conditional posteriors are

1. Draw $\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2)$
2. Draw $\mu|\sigma, y \sim N(\bar{y}, \sigma^2/n)$

Once again, generate 10000 draws from the posterior density and produce relevant summaries. Set the seed to 3.

`R` Tips:

- Data can be found in the file `hers.txt` under Datasets
- Useful functions: `rnorm()`, `rinvgamma()` from package `MCMCpack` with `shape` and `scale` defined accordingly

Conjugate Prior for Univariate Normal

The previous example can guide us in determining a joint conjugate prior for μ and σ^2

We showed that the joint posterior can be factored into two parts:

$$p(\sigma^2)p(\mu|\sigma^2)$$

Thus if our conjugate prior is of the same form, the posterior will be as well

I.B Multi-parameter Models

A convenient approach would be to take as conjugate the resulting marginal and conditional posteriors from the noninformative case

Recall that the marginal posterior of σ^2 was scaled inverse-χ^2 and the conditional posterior of μ was normal

Thus our conjugate prior could be made up of the following components:

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

I.B Multi-parameter Models

The resulting joint prior density is then

$$\begin{aligned} p(\mu, \sigma^2) &\propto p(\sigma^2) p(\mu|\sigma^2) \\ &= (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right) \\ &\times (\sigma^2)^{-1/2} \exp\left[-\frac{\kappa_0}{2\sigma^2}(\mu - \mu_0)^2\right] \\ &= (\sigma^2)^{-1/2} (\sigma^2)^{-(\nu_0/2+1)} \exp\left[-\frac{1}{2\sigma^2}\left\{\nu_0 \sigma_0^2 + \kappa_0(\mu - \mu_0)^2\right\}\right] \end{aligned}$$

which is known as the Normal-Inverse-χ^2 which is parameterized, in this case, by the location and scale of μ and the degrees of freedom and scale of σ^2

I.B Multi-parameter Models

Notes on the normal conjugate prior:

- μ and σ^2 are dependent
    - Since σ^2 appears in the prior for μ
- Thus a large σ^2 induces a high-variance prior distribution on μ
- As conjugate priors are largely used for convenience, we must note this and take care in specifying the hyper-parameters when using this conjugate prior

It's not all bad: prior beliefs on variance *should* be linked to prior beliefs on the mean to, at the very least, calibrate them to the scale of the measurements of y

I.B Multi-parameter Models

Example I.18

Let's now consider the posterior distribution that results from selecting the Normal-Inverse-χ^2 conjugate prior for the univariate normal model

Recall the form of the likelihood:

$$p(\mu, \sigma^2|y) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\bar{y}-\mu)^2\right\}\right]$$

I.B Multi-parameter Models

The posterior distribution is then

$$p(\mu, \sigma^2|y) \propto (\sigma^2)^{-1/2}(\sigma^2)^{-(\nu_0/2+1)} \exp\left[-\frac{1}{2\sigma^2}\left\{\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2\right\}\right]$$
$$\times (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\bar{y}-\mu)^2\right\}\right]$$

with some rearranging, it can be shown that this is also a Normal-Inverse-χ^2 (as it should be due to the conjugacy principle)

To see this, we must factor the posterior density into a conditional posterior of $\mu|\mu_0, \sigma_0^2, \kappa_0, y$, and a marginal posterior of $\sigma^2|\sigma_0^2, \kappa_0, y$

I.B Multi-parameter Models

Similar to the noninformative case, we can also find the conditional posterior distribution of $\mu|\sigma^2, y$

Example I.19

With some algebraic manipulations, it can be shown that

$$\mu|\sigma^2, y \sim N\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}\right)$$

which should remind us (in part) of the posterior for μ in the single-parameter model with known variance

I.B Multi-parameter Models

We can also get the marginal posterior of $\sigma^2 | y$

Example I.20

The marginal posterior will be a scaled inverse-χ^2 of the form

$$\sigma^2 | y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$$

where

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

I.B Multi-parameter Models

With this factorization, we could draw samples using the same two step procedure as before but updating the marginal and conditional posterior densities to reflect the use of the conjugate prior:

1. Draw $\sigma^2 | y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$
2. Draw $\mu | \sigma, y \sim N\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}\right)$

Once again, we use the *marginal-conditional* approach where for step 2, we use the simulated values of σ^2

We can also obtain the marginal posterior of $\mu | y$ here

I.B Multi-parameter Models

Using an analogous approach to the one we used in the noninformative case, we can get the marginal posterior of $\mu|y$ for the conjugate case

Example I.21

The marginal posterior of $\mu|y$ will be a t-distribution with degrees of freedom ν_n, location μ_n, and scale σ_n^2/κ_n where ν_n and σ_n^2 are as defined above and

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

I.B Multi-parameter Models

Multinomial distribution

For random variables whose outcomes fall into one of k categories. This is a multivariate generalization of the binomial distribution such that the marginal distribution of a single θ_i is binomial.

Notation: $\theta \sim Multinom(n; p_1, \ldots, p_k)$ $\qquad$ $E(\theta_j) = np_j$

PMF: $p(\theta) = \binom{n}{\theta_1\ \theta_2\ \ldots\ \theta_k} p_1^{\theta_1} \cdots p_k^{\theta_k}$ $\qquad$ $Var(\theta_j) = np_j(1 - p_j)$

Notes: $\theta = [\ \theta_1 \quad \theta_2 \quad \cdots \quad \theta_k\ ]'$, $\theta_j = 0, 1, 2, \ldots, n$, and $\sum_j \theta_j = n$

Useful `R`-functions: `rmultinom`, `dmultinom`

I.B Multi-parameter Models

Generalizing the binomial to more than two categories gives us the multinomial

Let y be a vector of counts of the number of observations of each of the k outcomes take from a multinomial sample

The likelihood is then:

$$\mathcal{L}(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}$$

where $\sum_j \theta_j = 1$, i.e. the sum of the probabilities is 1

I.B Multi-parameter Models

Question: What was the conjugate prior family for the binomial?

I.B Multi-parameter Models

Question: What was the conjugate prior family for the binomial?

The beta distribution

I.B Multi-parameter Models

Question: What was the conjugate prior family for the binomial?

The beta distribution

Since the multinomial is a multivariate generalization of the binomial, perhaps there is a multivariate generalization of the the beta that will serve as conjugate

I.B Multi-parameter Models

Question: What was the conjugate prior family for the binomial?

The beta distribution

Since the multinomial is a multivariate generalization of the binomial, perhaps there is a multivariate generalization of the the beta that will serve as conjugate

Question: Does such a distribution exist?

I.B Multi-parameter Models

Dirichlet distribution

The multivariate generalization of the beta distribution. The marginal distributions of a single parameter is Beta.

Notation: $\theta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$

$E(\theta_j) = \frac{\alpha_j}{\alpha_0}$

Density: $p(\theta) = \frac{\Gamma(\sum_j^k \alpha_j)}{\prod_j^k \Gamma(\alpha_j)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$

$Var(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$

Notes: $\theta = [\ \theta_1 \quad \theta_2 \quad \cdots \quad \theta_k\ ]'$, $\theta_1, \ldots, \theta_k \geqslant 0$, $\sum_{j=1}^k \theta_j = 1$, and $\alpha_0 \equiv \sum_{j=1}^k \alpha_j$

Useful `R`-functions: `rdirichlet` from the package `MCMCpack`

I.B Multi-parameter Models

Example I.22

Suppose we want to study the polls from 2017 French Presidential Election. There were many parties, but suppose we restrict to three candidates: Le Pen (National Front), Fillon (the Republicans), and Marcon (En Marche!). We'll also consider a fourth category: other (including all other candidates as well as those who abstained).

Let's build a multinomial model for this data

Also determine a noninformative prior and find the form of the posterior

I.B Multi-parameter Models

Example I.R7

Ifop-Fiducial conducted a poll between January 29th and February 1st of 2017 from a sample of 1409 French voters. The data is summarized in the table below.

Party	Nat. Front	Republicans	En Marche!	Other
Candidate	LePen	Fillon	Macron	Others/?
Count	338	296	282	493

Find posterior summaries for the model parameters and determine the posterior probability that Macron would get at least 24.01% of the vote in the first round based on 10000 draws.

`R` Tips:

- `rdirichlet()` returns a $B \times k$ matrix, to summarize, use `apply()` along with model summaries, set the seed to 2584

I.B Multi-parameter Models

With the Dirichlet, we can easily model data arising from a multinomial likelihood

When our data is continuous, however, we often rely (for better or for worse) on the multivariate normal

This distribution is incredibly useful and can arise in a wide range of settings for continuous (and occasionally discrete) likelihoods

I.B Multi-parameter Models

Multivariate normal distribution

The generalization of the normal distribution to multiple outcomes. It allows for the modeling of multiple normally distributed random variables that can be correlated based on a covariance matrix. Let $\boldsymbol{\theta}$ be a vector of normally distributed r.v.'s and let $\boldsymbol{\mu}$ be a vector of means, i.e. $\boldsymbol{\theta} = [\ \theta_1 \quad \theta_2 \quad \cdots \quad \theta_D\ ]'$ and $\boldsymbol{\mu} = [\ \mu_1 \quad \mu_2 \quad \cdots \quad \mu_D\ ]'$.

Notation: $\boldsymbol{\theta} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is $D \times 1$ and $\boldsymbol{\Sigma}$ is $D \times D$

PDF: $p(\boldsymbol{\theta}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]$

$E(\boldsymbol{\theta}) = \boldsymbol{\mu}$ and $Var(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$ or $Cov(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$

Useful `R`-functions: `rmvnorm` and `dmvnorm` from package `mvtnorm`

I.B Multi-parameter Models

Some properties of the multivariate normal:

1. The marginal distribution of a single component of θ is also normal:
$$\theta_d \sim N(\mu_d, \sigma_d^2)$$
2. The marginal distribution of any sub-vector of θ is multivariate normal. Let $\theta^* = [\ \theta_d \quad \theta_{d+1}\ ]'$, then
$$\theta^* \sim MVN(\boldsymbol{\mu}*, \boldsymbol{\Sigma}^*)$$
for $\boldsymbol{\mu}^* = [\ \mu_d \quad \mu_{d+1}\ ]'$ and corresponding 2 × 2 covariance matrix $\boldsymbol{\Sigma}^*$
3. The conditional distribution of $\theta_d|\theta_d + 1$ is also normally distributed

I.B Multi-parameter Models

In general, the covariance matrix has the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{pmatrix}.$$

where σ_{ij} is the covariance between θ_i and θ_j and σ_d^2 is the variance of θ_d.

Other structures may be imposed, but this is the most general

In particular, independence may be something we want to build in to model

I.B Multi-parameter Models

If we assume independence among the elements of θ, but heterogeneity in the variances, we get a covariance matrix that has the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{pmatrix}.$$

where the off-diagonals are all zero but the diagonal is made up of component specific variances

As before, $Var(\theta_d) = \sigma_d^2$ but now $Cov(\theta_j, \theta_i) = 0 \; \forall \; i \neq j$

In the multivariate normal case, $Cov(\theta_j, \theta_i) = 0 \iff \theta_i \perp \theta_j$

I.B Multi-parameter Models

A common assumption when using the multivariate normal is that there is a common variance, i.e. the variances are homogeneous. This structure has the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma^2 \mathbf{I}_{d \times d}$$

where $\mathbf{I}_{d \times d}$ is the $d \times d$ identity matrix, thus, $Var(\theta_d) = \sigma^2 \; \forall \; d$ and $Cov(\theta_j, \theta_i) = 0 \; \forall \; i \neq j$

This is one of the main assumptions in linear regression

I.B Multi-parameter Models

Linear regression is one of the most widely used statistical tools

There are several ways to obtain estimates from a regression model in the Bayesian context

In this unit, we will consider one of the simpler approaches that uses the marginal-conditional approach

First, a review of ideas from regression

I.B Multi-parameter Models

Let y denote the *response* or *outcome* variable which is the primary quantity of interest and assumed, for now, to be continuous

Further, let $x = (x_1, \ldots, x_k)$ denote the explanatory variables that may be discrete or continuous

Often there is one x_j that we are primarily interested in relating to y while controlling for the remaining the variables

I.B Introduction to Regression Models

Typically we have a set of observations on our subjects which we denote with a subscript i, thus

$$y_i \text{ and } x_{i1}, \ldots, x_{ik}$$

denote subject i's observed outcome and set of covariates

In general, we will let $y = [\ y_1 \quad \cdots \quad y_n\ ]'$, the $n \times 1$ vector of outcomes, and X denote the $n \times k$ matrix of covariates (or predictors or explanatory variables)

I.B Multi-parameter Models

Least-squares Review

Suppose we are interested in fitting the model

$$y = X\boldsymbol{\beta} + \epsilon$$

where $\boldsymbol{\beta} = [\begin{array}{cccc} \beta_0 & \beta_1 & \cdots & y_k \end{array}]'$ and ϵ is a vector of mean zero random errors with finite variance

Least-squares estimation minimizes the following score equation

$$Q = (y - X\boldsymbol{\beta})'\Sigma^{-1}(y - X\boldsymbol{\beta})$$

where $var(\epsilon) = \Sigma$

I.B Multi-parameter Models

Least-squares Review (cont.)

If we assume the errors are iid, that is the ϵ_i come from the same distribution with common variance σ^2, then $\Sigma = \sigma^2 I_{n \times n}$ and the least-squares equation reduces to

$$Q = \frac{1}{\sigma^2}(y - X\boldsymbol{\beta})'(y - X\boldsymbol{\beta})$$

Minimizing and solving for $\boldsymbol{\beta}$ gives the least squares estimate of the regression line

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y$$

I.B Multi-parameter Models

Normal Regression

The least-squares approach requires very few assumptions, namely the errors have mean zero and the variance is finite (and even mean zero is flexible)

But it is very common to assume the errors are in fact normal, thus

$$\epsilon \sim MVN\left(0, \sigma^2 I_{n\times n}\right)$$

This assumption, in turn, means that we assume

$$y \sim MVN\left(X\boldsymbol{\beta}, \sigma^2 I_{n\times n}\right)$$

I.B Multi-parameter Models

A note on X:

X is often called the *design matrix* and the leading column is typically a vector of ones:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

The first column corresponds to the intercept which, while usually estimated, is not often the focus on inference

I.B Multi-parameter Models

Goal of Estimation

In the regression setting, we usually consider X and $\boldsymbol{\beta}$ to be 'fixed' components, thus the only component of the RHS of the model that is random is ϵ

Our goal is to model the mean of y given X and $\boldsymbol{\beta}$, i.e. we want to model

$$E(y|\boldsymbol{\beta}, X) = X\boldsymbol{\beta}$$

since $E(\epsilon) = 0$

I.B Multi-parameter Models

Bayesian Analysis of Classical Regression

As noted above, we begin by assuming the likelihood of y depends on the covariates X, the unknown coefficients $\boldsymbol{\beta}$, and the unknown variance σ^2, thus

$$y \sim MVN\left(X\boldsymbol{\beta}, \sigma^2 I_{n\times n}\right)$$

Question: what prior(s) do we place on $\boldsymbol{\beta}$ and σ^2?

Bayesian Analysis of Classical Regression

As noted above, we begin by assuming the likelihood of y depends on the covariates X, the unknown coefficients $\boldsymbol{\beta}$, and the unknown variance σ^2, thus

$$y \sim MVN\left(X\boldsymbol{\beta}, \sigma^2 I_{n\times n}\right)$$

Question: what prior(s) do we place on $\boldsymbol{\beta}$ and σ^2?

Typically, we assume the non-informative prior:

I.B Multi-parameter Models

Bayesian Analysis of Classical Regression

As noted above, we begin by assuming the likelihood of y depends on the covariates X, the unknown coefficients $\boldsymbol{\beta}$, and the unknown variance σ^2, thus

$$y \sim MVN\left(X\boldsymbol{\beta}, \sigma^2 I_{n\times n}\right)$$

Question: what prior(s) do we place on $\boldsymbol{\beta}$ and σ^2?

Typically, we assume the non-informative prior:

$$p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$$

Note: this is Jeffreys' prior

I.B Multi-parameter Models

Notes on $p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$:

- When there are many data points and only a few parameters, the noninformative prior distribution is useful and gives acceptable results
- However, when n is small or k is large, the likelihood is less sharply peaked and so prior distributions and hierarchical models are more important

If time allows, we will consider some of these issues (particularly if k is large)

I.B Multi-parameter Models

Normal Model Posterior

We can use the same approach here as in Part I for the univariate normal model where we factor the density into a conditional and (nearly) marginal posterior

Thus our posterior has the form

$$p(\boldsymbol{\beta}, \sigma^2|y) \propto p(\boldsymbol{\beta}|\sigma^2, y)p(\sigma^2|y)$$

Example I.23

Write out the model and derive the conditionals

I.B Multi-parameter Models

Conditional Posterior of $\boldsymbol{\beta}$ given σ^2

The conditional posterior distribution of the vector $\boldsymbol{\beta}$ given σ^2 is a quadratic form in $\boldsymbol{\beta}$ and thus normal

$$\boldsymbol{\beta}|\sigma^2, y \sim MVN\left(\hat{\boldsymbol{\beta}}, V_{\boldsymbol{\beta}}\sigma^2\right)$$

where $\hat{\boldsymbol{\beta}}$ and $V_{\boldsymbol{\beta}}$ have the form

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'y \\ V_{\boldsymbol{\beta}} &= (X'X)^{-1}\end{aligned}$$

I.B Multi-parameter Models

Marginal Posterior of σ^2

The marginal posterior can be written as we did in Part I, as the ratio of the full posterior to the conditional posterior of $\boldsymbol{\beta}$:

$$p(\sigma^2|y) = \frac{p(\boldsymbol{\beta}, \sigma^2|y)}{p(\boldsymbol{\beta}|\sigma^2, y)}$$

which can be seen to have a scaled inverse-χ^2 form:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-k, s^2)$$

for s^2 given by

$$s^2 = \frac{1}{n-k}\left(y - X\hat{\boldsymbol{\beta}}\right)'\left(y - X\hat{\boldsymbol{\beta}}\right)$$

I.B Multi-parameter Models

Marginal Posterior of $\boldsymbol{\beta}$

It can be shown that, after averaging over σ^2, the marginal posterior of $\boldsymbol{\beta}|y$ is multivariate t with $n-k$ degrees of freedom

However, we rarely use this fact in practice when drawing inferences by simulation since to characterize the joint posterior we can draw simulations of σ and then $\boldsymbol{\beta}|\sigma$

In other words, we use the *marginal-conditional* approach

I.B Multi-parameter Models

Example I.R8

Download the dataset `hersreg.txt` from the data folder. Build a Bayesian linear regression model using change in total cholesterol, `chtchol` as the outcome and `treatment` as the primary covariate of interest. Also control for baseline systolic blood pressure, `sbp`, as well as statin use, `statins`.

Implement a *marginal-conditional* sampler to run this model with $\mathrm{B} = 10000$, use `rmvnorm()` from the `mvtnorm` package, and set the seed to 90210

I.B Multi-parameter Models

Returning to the non-regression case, we can also describe a general multivariate normal

As with the univariate case, we begin by examining the multivariate normal where the covariance is known and fixed

Example I.24

Suppose $\mathbf{y} = [\begin{array}{cccc} y_1 & y_2 & \cdots & y_d \end{array}]'$ is a vector of normally distributed random variables with unknown mean, $\boldsymbol{\mu}$, and fixed variance, $\boldsymbol{\Sigma}$. Determine the form of the likelihood.

I.B Multi-parameter Models

Given the quadratic form of the likelihood, having a prior that is also of quadratic form would make sense

Thus the conjugate prior is $\boldsymbol{\mu} \sim \mathrm{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ which has the form

$$\pi(\boldsymbol{\mu}) \propto |\boldsymbol{\Lambda}_0|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\Lambda}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right]$$

What is the form of the resulting posterior distribution?

For the multivariate normal, we can also determine conditional and marginal distributions of the subcomponents of the vector $\boldsymbol{\mu}$

The conjugate prior can be used to make an informative or weakly informative prior

A noninformative prior is $\pi(\boldsymbol{\mu}) \propto c$ for some constant c

Question: What happens when $\boldsymbol{\Sigma}$ is also unknown?

I.B Multi-parameter Models

Wishart distribution

A multivariate generalization of a gamma distribution, can be used to model the inverse of a covariance matrix

Notation: $\mathcal{W} \sim \mathcal{W}_\nu(\mathcal{S})$, with ν degrees of freedom and $k \times k$ scale matrix $\mathcal{S}$

PDF: $p(\mathcal{W}) \propto |S|^{-\nu/2}|\mathcal{W}|^{(\nu-k-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\mathcal{S}^{-1}\mathcal{W}\right)\right]$

$E(\mathcal{W}) = \nu \mathcal{S}$

Useful R-functions: `rwish` and `dwish` from package `MCMCpack`

I.B Multi-parameter Models

Inverse-Wishart distribution

A multivariate generalization of a scaled inverse-χ^2 distribution, can be used to model the covariance matrix

Notation: $W \sim \mathrm{IW}_\nu(S^{-1})$, with ν degrees of freedom and $k \times k$ scale matrix S

PDF: $p(W) \propto |s|^{\nu/2} |W|^{-(\nu+k+1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(SW^{-1}\right)\right]$

$E(W) = (\nu - k - 1)^{-1} S$

Useful R-functions: `riwish` and `diwish` from package `MCMCpack`

I.B Multi-parameter Models

Marginalization Property

Let W be a random matrix arising from an inverse-Wishart, i.e. $W \sim \mathrm{IW}_{\nu}(S^{-1})$. Then any principal sub-matrix of W is also inverse-Wishart.

This is known as the marginalization property of the inverse-Wishart

It can be useful for sampling covariance matrices when, for instance, they are assumed to be block-diagonal

Example I.25

Recall that the conjugate distribution for the univariate normal case with unknown mean and variance is the normal-inverse-χ^2. Thus for the multivariate normal, it would make sense that the conjugate prior be the multivariate realization of this, i.e. the multivariate-normal-inverse-Wishart or just the normal-inverse-Wishart.

I.B Multi-parameter Models

Suppose $y = [\ y_1 \quad y_2 \quad \cdots \quad y_d\ ]'$ is a vector of normally distributed r.v.'s with unknown mean $\boldsymbol{\mu}$ and unknown variance $\boldsymbol{\Sigma}$.

The likelihood then has the form:

$$\mathcal{L}(y|\boldsymbol{\mu}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(y-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(y-\boldsymbol{\mu})\right]$$

Thus, the conjugate prior has the form

$$\boldsymbol{\Sigma} \sim IW_{\nu_0}\left(\boldsymbol{\Lambda}_0^{-1}\right)$$
$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim MVN\left(\boldsymbol{\mu}_0, \kappa_0^{-1}\boldsymbol{\Sigma}\right)$$

I.B Multi-parameter Models

The resulting joint prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-((\nu_0+d)/2+1)} \times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}\right) - \frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right]$$

Multiplying this prior density by the likelihood will result in the posterior having a normal-inverse-Wishart parametric form

Further, it can be shown that other results from the univariate case generalize to the multivariate case, for example the marginal posterior of $\boldsymbol{\mu}$ is a multivariate t

I.B Multi-parameter Models

Noninformative priors

Suppose we wish to implement a non-informative prior for the multivariate normal model with unknown mean and variance

Inverse-Wishart with $d + 1$ df

We can alter the conjugate prior to give a prior places a uniform prior on each correlation in $\boldsymbol{\Sigma}$, marginally (however the joint is not uniform)

Such a prior has the form

$$\boldsymbol{\Sigma} \sim IW_{d+1}\left(\mathbf{I}_{d \times d}\right)$$

along with a multivariate normal on $\boldsymbol{\mu}|\boldsymbol{\Sigma}$

I.B Multi-parameter Models

Jeffreys' Prior

The multivariate Jeffreys' prior has the form

$$\pi(\mu, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(d+1)/2}$$

This prior also arises by taking the conjugate prior and letting $\kappa_0 \to 0$, $\nu_0 \to -1$, and $|\boldsymbol{\Lambda}_0| \to 0$

Using this prior results in a posterior of the form:

$$\boldsymbol{\Sigma}|y \sim IW_{n-1}(S^{-1})$$
$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, y \sim MVN(\bar{y}, n^{-1}\boldsymbol{\Sigma})$$

It can also be shown that

$$\boldsymbol{\mu}|y \sim MVT_{n-d}[\bar{y}, (n(n-d))^{-1}S]$$

Unit C: Normal Approximation

BDA Chapters 4 and 16

I.C Normal Approximation

In many contexts, the Bayesian model based on the noninformative prior results in a model similar to that of the Frequentist approach

Example I.26

For the normal model, when the mean and variance are unknown, the marginal posterior of the mean follows a t distribution when using the noninformative prior

In many of the models we have discussed, the influence of the prior decreases with an increased sample size, n

I.C Normal Approximation

Influence of n

Note that this does not mean that Bayesian models *require* large n, as is the case for many Frequentist approaches, only that a large n mitigates the influence of the prior

In this way, the choice of prior becomes almost negligible when n is sufficiently large

But a large n also provides us with an alternative way to sample from the posterior via a normal approximation

I.C Normal Approximation

Normal Approximation to the Posterior

If $p(\theta|y)$ is unimodal and roughly symmetric, we can approximate it with a normal distribution

Specifically, $\log\left[p(\theta|y)\right]$ is approximated by a quadratic function

Here we consider a quadratic approximation to the log-posterior centered at the posterior mode

I.C Normal Approximation

Recall from Calc II:

Taylor Series Expansion

For a function $f(x)$, the Taylor Series Expansion about a is given by

$$
\begin{aligned}
f(x) &= \sum_{i=0}^{\infty} \frac{d^i}{da^i} f(a) \frac{(x-a)^i}{i!} \\
&= f(a) + \frac{d}{da} f(a)(x-a) + \frac{d^2}{da^2} f(a) \frac{(x-a)^2}{2!} + \cdots
\end{aligned}
$$

I.C Normal Approximation

Applying the expansion to $\log\left[p(\theta|y)\right]$ and allowing for θ to potentially be a vector, we get

$$\log\left[p(\theta|y)\right] = \log\left[p(\hat{\theta}|y)\right] + \frac{1}{2}(\theta - \hat{\theta})' \left[\frac{d^2}{d\theta^2}\log\left[p(\theta|y)\right]\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots$$

where $\hat{\theta}$ is the posterior mode (hence the linear term is zero as the log-posterior has zero derivative at its mode)

When θ is close to $\hat{\theta}$ and n is large, the higher order terms fade in importance relative to the quadratic term

I.C Normal Approximation

$$\log\left[p(\theta|y)\right] = \log\left[p(\hat{\theta}|y)\right] + \frac{1}{2}(\theta-\hat{\theta})'\left[\frac{d^2}{d\theta^2}\log\left[p(\theta|y)\right]\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + \cdots$$

Considering the above equation, note that the first term is a constant

Next note that the second term is proportional to the log of a normal density

This forms the basis of the approximation

I.C Normal Approximation

For a large enough n and $\hat{\theta}$ in the interior of the parameter space,

$$p(\theta|y) \approx N\left(\hat{\theta}, [I(\hat{\theta})]^{-1}\right)$$

where $I(\theta)$ is the observed information,

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

Note: if the mode is in the interior of the parameter space, then $I(\hat{\theta})$ will be positive definite

I.C Normal Approximation

Let's consider a simple example:

Example I.27

Let $y_i \sim Pois(\lambda)$ and assume the noninformative prior $\pi(\lambda) \propto \lambda^{-1}$. Find the posterior distribution, the log-posterior, posterior mode, and the resulting information.

Question: What form of the normal distribution approximates the posterior resulting from this model?

I.C Normal Approximation

Note that the posterior has a Gamma form:

$$p(\lambda|y) \propto \lambda^{\sum y_i - 1} e^{-n\lambda}$$

Differentiating the log-posterior, we obtain the posterior mode and inverse-information as

$$\hat{\lambda} = \frac{\sum y_i - 1}{n} \text{ and } \left[I(\hat{\lambda})\right]^{-1} = \frac{\sum y_i - 1}{n^2}$$

Thus $p(\lambda|y)$ can be approximated by a normal with mean $\hat{\lambda}$ and variance $\left[I(\hat{\lambda})\right]^{-1}$

I.C Normal Approximation

Example I.R9

Consider data on the number of reported cases of Lyme disease (a CDC notifiable disease) in each county of Minnesota during 2010. The data can be found in the file `lyme.txt` and, with some assumptions, can be considered Poisson. Use this data to generate posterior draws using first the normal approximation and then the closed form Gamma. Take 1000 samples, setting the seed to 217.

Questions:

1. By how much do the posterior estimates and intervals differ?
2. How similar are the posterior densities?

Increase the number of samples to 10000, compare again

I.C Normal Approximation

The normal approximation is, of course, more useful when there are multiple unknown parameters and the posterior does not have a known form

For example, generalized linear models can be fit in the Bayesian context using this approach

I.C Normal Approximation

Often we are interested in regression models that have categorical outcomes rather than normally distributed outcomes

In the frequentist setting, this is accomplished via a generalized linear model or GLM with appropriate link function

We will now briefly review GLM theory from the frequentist perspective before considering Bayesian approaches to modeling categorical outcomes

I.C Normal Approximation

(Agresti, Sections 4.1 and 4.4)

Responses: $y_1, \ldots, y_n$ (independent random variables)

Covariates: $x_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$

- (throughout, vectors are column vectors)
- If the model includes a constant term, then $x_{i1} \equiv 1$

A *generalized linear model*

- Specifies a parametric family for the conditional distribution of the y_i given the x_i
- The model is specified through
  1) a random component,
  2) a systematic component, and
  3) a link function relating the two.

I.C Normal Approximation

1) *Random Component*: specifies an extended exponential family of distributions for the Y_i, with a pdf/pmf of the form

$$f(y_i; \theta_i, \phi) = \exp[\{y_i\theta_i - b(\theta_i)\}/a_i(\phi) + c(y_i, \phi)] \qquad (1)$$

- ϕ can be a known constant or an unknown parameter
- If ϕ is known, (1) is a regular exponential family
- θ_i is the *canonical parameter*; natural parameter of a regular exponential family in some cases
- In general, a_i (and also c) can vary with i

I.C Normal Approximation

2) The *systematic component* is specified through a *linear predictor*

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = x_i'\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and the β_j are unknown parameters

- As with the normal model, x_i can include
  - Polynomial terms or transformations of the covariate values
  - Indicator variables for subsets defined by covariates
  - Interaction terms and other functions of the observed covariates
  - Smoothed effects
  - Etc.

I.C Normal Approximation

3) The *link function*, $g(\cdot)$, specifies the relationship between η_i and

$$\mu_i = \mathsf{E}(Y_i) = \mathsf{E}(Y_i|\mathbf{x}_i)$$

through

$$\eta_i = g(\mu_i).$$

Link functions

- Are strictly monotone
- Have well defined inverse functions $h(\cdot) = g^{-1}(\cdot)$ satisfying
  - $h\{g(\mu_i)\} = \mu_i$ and $g\{h(\eta_i)\} = \eta_i$
  - So, $h(\eta_i) = \mu_i$

I.C Normal Approximation

Canonical Link

Notice that linear regression, logistic regression, and Poisson regression with a log link all use

$$g(\mu_i) = \theta_i$$

That is, these models correspond to setting the link function equal to the canonical parameter in the assumed natural exponential distribution for the response. This link has a special place in GLM theory, and is called the *canonical link function*.

I.C Normal Approximation

The GLM approach forms a unified model for conducting regression when y_i is normal, Bernoulli, binomial, multinomial, poisson, exponential, etc.

In the Bayesian context, it's less unified as there are a number of different ways to model categorical data (of course this can be to our advantage)

The three approaches we will consider are

- Normal Approximation based models (Part I)
- Latent Variable based models (Part II)
- Metropolis-Hastings based models (Part II)

I.C Normal Approximation

Normal Approximation

For simple computations, it can be convenient in the GLM setting to approximate the likelihood by a normal distribution in β conditional on the dispersion parameter ϕ

For each y_i, we construct a *pseudodatum* z_i and a *pseudovariance* σ_i^2 so that the GLM likelihood $\mathcal{L}(y_i|x_i'\beta, \phi)$ is approximated by the normal likelihood $N(z_i, \sigma_i^2)$

The method aims to approximate the GLM by a linear model

I.C Normal Approximation

In general, a standard way to determine z_i and σ_i^2 for the approximation is to match the first and second order terms of the Taylor expansion of the log-likelihood, $\ell(y_i|x_i'\beta, \phi)$ centered about $\hat{\eta}_i = x_i'\hat{\beta}$:

$$z_i = \hat{\eta}_i - \frac{\partial}{\partial \eta}\ell(y_i|x_i'\beta, \phi) \Bigg/ \frac{\partial^2}{\partial \eta^2}\ell(y_i|x_i'\beta, \phi)$$

$$\sigma_i^2 = -\left[\frac{\partial^2}{\partial \eta^2}\ell(y_i|x_i'\beta, \phi)\right]^{-1}$$

The posterior mode can then be found using iterative weighted linear regression or, equivalent, Fisher's scoring (Newton-Raphson)

I.C Normal Approximation

Once the posterior mode, $\hat{\beta}$ and $\hat{\phi}$ has been reached, one can approximate the conditional posterior distribution of β given $\hat{\phi}$

Approximate Normal Posterior

The posterior of β can then be sampled from

$$\beta|\hat{\phi}, y \sim N(\hat{\beta}, V_{\beta})$$

where V_{β} is the last working variance from iterative solution to the posterior mode

I.C Normal Approximation

Logistic Model using Normal Approximation

In practice, the pseudodata formulation and Fisher scoring approach are what is used in the frequentist estimation of GLMs

Thus $\hat{\beta}$ and V_{β} can first be obtained from a standard GLM method

Then the desired number of samples can be taken from $\mathrm{N}(\hat{\beta}, V_{\beta})$ with inference conducted as usual

I.C Normal Approximation

Code to implement the normal approximation approach to logistic regression can be found on the website

Example I.R10

Consider a dataset detailing Old Faithful eruptions found in the file `oldFaithful.txt`. The outcome variable, `etime`, is code 1 if the eruption lasted longer than three minutes, 0 otherwise. We wish to predict `etime` with the waiting time to said eruption, `waiting`.

Note: this approach involves direct sampling, though the posterior is based on an approximation

I.C Normal Approximation

Poisson Regression using Normal Approximation

Similar to the binary case, if y_i is Poisson, we can easily implement the normal approximation approach to draw samples from the posterior distribution of β after using Fisher's scoring

Once again, the posterior of β is $N(\hat{\beta}, V_\beta)$ given the posterior mode and working covariance from the last iteration of Fisher's scoring

I.C Normal Approximation

Code to implement the normal approximation approach to Poisson regression can be found on the website

Example I.R11

Consider a study of seizures in patients with epilepsy who were randomized to two treatment groups: placebo vs. active treatment with progabide. The number of seizures over a two week period was measured for each subjects for eight weeks. While this study was longitudinal in nature, we focus on the last two-week period to see if by the end of the study the treatment had decreased seizure activity. The data is in the file `seizures.txt`, the outcome is the variable `count` with predictors treatment, `trt`, age, baseline number of seizures, `base`.

Units B &C Summary

Take-home message: elementary modeling and computation

Despite the lack of easily derived closed form distribution for multi-parameter models, we see that we have some tools available to us to sample these parameters (and more on the way)

So far, we can use

- Direct sampling (if closed form is available)
- Marginal-conditional approach
- Full marginals (obtained via integration)
- Normal approximations

Units B &C Summary

Strategy for Computation of Simple Bayesian Posteriors

1. Determine $\mathcal{L}(y|\theta)$ ignoring components that are free of θ
2. Determine the posterior $p(\theta|y) \propto \mathcal{L}(y|\theta)\pi(\theta)$
   - Include prior information if relevant, use a weakly informative prior, or set the prior to constant and update later
3. Determine sampling technique best suited to problem
4. Draw simulations $\theta^1, \ldots, \theta^B$ from the posterior
   - For non-conjugate models, this step can be difficult
5. Use sample draws to compute posterior summaries
6. (If desired) Compute predictive quantities, simulate $\tilde{y}^1, \ldots, \tilde{y}^B$ by drawing $\tilde{y}^b$ from $p(\tilde{y}|\theta^b)$.

Unit C: Hyper Priors and Hierarchical Models

BDA Chapter 5

I.D Hyper Priors and Hierarchical Models

Thus far, most of our models have consisted of prior distributions that either have fixed parameters, as in conjugacy, or were non-informative, often a constant wrt the parameter

However, a prior distribution could contain additional parameters that are not part of the likelihood

Example I.28

For example, suppose we wish to model the data y through the likelihood $\mathcal{L}(y|\theta)$, if we place the prior $\pi(\theta|\phi)$ on θ, we have introduced a new parameter, a *hyper-parameter*

I.D Hyper Priors and Hierarchical Models

Thus the posterior has the form

$$p(\phi, \theta|y) \propto \mathcal{L}(y|\theta)\pi(\theta|\phi)$$

Before we set $\phi = \phi_0$, some fixed quantity, but if we allow ϕ to vary, we could place a prior on it, i.e. a *hyper-prior* on our *hyper-parameter*

In other words the distribution of the data depends only on θ and the hyper-parameter ϕ only affects y through θ

The final model could then be

$$p(\phi, \theta|y) \propto \mathcal{L}(y|\theta)\pi(\theta|\phi)\pi(\phi)$$

I.D Hyper Priors and Hierarchical Models

In order to model the uncertainty in ϕ, we must place a prior on it

Hyperprior

If we know little about ϕ, we can assign a diffuse prior density

In most real problems, we should have enough substantive knowledge about ϕ to at least constrain the hyper-parameters into a finite region

A relatively simple, noninformative prior is useful to start and if the variation is too great in the posterior, we can add more information

I.D Hyper Priors and Hierarchical Models

This is an of a simple hierarchical model

To build such a model, we perform the following three steps:

1 Write the joint posterior density as $\mathcal{L}(y|\theta)\pi(\theta|\phi)\pi(\phi)$

2 Analytically determine the conditional posterior of $\theta|\phi, y$

3 Estimate ϕ by obtaining its marginal posterior

Let's consider each step

I.D Hyper Priors and Hierarchical Models

1 Write the joint posterior

This step is immediate given our choice of the likelihood, prior, and hyper-prior

2 Find conditional posterior

If we select a conjugate prior for θ, then this step is easy since as the conditional of $\theta|y, \phi$ will be known, albeit dependent upon ϕ

I.D Hyper Priors and Hierarchical Models

3 Obtain the marginal posterior

We can do this by brute force

$$p(\phi|y) = \int p(\theta, \phi|y)d\theta$$

However, for many models we can compute this using conditional probability:

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}$$

Note: the numerator is the joint posterior and the denominator is the posterior for θ if ϕ were known

Also note that we must be careful with the proportionality constant to ensure it is actually a constant

I.D Hyper Priors and Hierarchical Models

The steps to modeling are then

1. Take a single draw of the hyper-parameter ϕ from the posterior marginal
2. Take a single draw from the conditional posterior $p(\theta|\phi, y)$ given the drawn value of ϕ
3. Draw predictive values for $\tilde{y}$ (if desired)

Repeat these steps B times

Summarize parameters using posterior inference as before

I.D Hyper Priors and Hierarchical Models

We can extend this idea to the case where we have many parameters of interest that can be regarded as related or connected in some way

Example I.29

Patients being treated for pancreatic cancer at center j may have a probability of one-year survival of θ_j. Further, it may be reasonable to expect estimates of the θ_j's to be related to each other.

Key here is that the observed data, y_{ij}, is indexed by i within units and j within groups which can still be used to model θ_j.

This is an example of *hierarchical data*

I.D Hyper Priors and Hierarchical Models

In practice, nonhierarchical models are usually inappropriate for hierarchical data

It's a Goldilocks problem:

I.D Hyper Priors and Hierarchical Models

In practice, nonhierarchical models are usually inappropriate for hierarchical data

It's a Goldilocks problem:

Too few

Models cannot fit large datasets accurately

I.D Hyper Priors and Hierarchical Models

In practice, nonhierarchical models are usually inappropriate for hierarchical data

It's a Goldilocks problem:

Too few

Models cannot fit large datasets accurately

Too many

Models can overfit such data

I.D Hyper Priors and Hierarchical Models

In practice, nonhierarchical models are usually inappropriate for hierarchical data

It's a Goldilocks problem:

Too few

Models cannot fit large datasets accurately

Too many

Models can overfit such data

Just right

Hierarchical models fit the data well while using a population distribution to structure some dependencies into the parameters

I.D Hyper Priors and Hierarchical Models

Consider a set of experimenst $j = 1, \ldots, J$ in which experiment j has data $\mathbf{y}_j$ and parameter θ_j with likelihood $\mathcal{L}(\mathbf{y}_j|\theta_j)$

Some of the parameters may overlap:

Example I.30

$\mathbf{y}_j$ may be a sample of observations from a normal distribution with mean $\boldsymbol{\mu}_j$ and common variance σ^2

We need the idea of *exchangeability* to create a joint probability model

I.D Hyper Priors and Hierarchical Models

Exchangeability

If no information, other than the data, is available to distinguish any of the θ_j's from any others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution.

This symmetry is represented probabilistically by exchangeability: the parameters $(\theta_1, \ldots, \theta_J)$ are exchangeable in their joint distribution if $p(\theta_1, \ldots, \theta_J)$ is invariante to permutations of the indexes $(1, \ldots, J)$.

I.D Hyper Priors and Hierarchical Models

Exchangeability is a concept we have already encountered when constructing iid models for our data

As Gelman notes, “ignorance implies exchangeability”

Example I.31

Consider rolling a die. Initially, we should assign equal probabilities to all six outcomes. This is assuming exchangeability. After some observation, we may notice imperfections which might make us favor one outcome over the us and thus eliminate the symmetry.

I.D Hyper Priors and Hierarchical Models

Basic form of an exchangeable distribution

Each of the parameters θ_j is an independent sample from a prior distribution governed by some unknown parameter vector $\boldsymbol{\phi}$, thus

$$p(\theta|\boldsymbol{\phi}) = \prod_{j=1}^{J} p(\theta_j|\boldsymbol{\phi})$$

Since $\boldsymbol{\phi}$ is usually unknown, our distribution for θ must average over our uncertainty in $\boldsymbol{\phi}$:

$$p(\theta) = \int \left[\prod_{j=1}^{J} p(\theta_j|\boldsymbol{\phi})\right] p(\boldsymbol{\phi}) d\boldsymbol{\phi}$$

I.D Hyper Priors and Hierarchical Models

This form, the mixture of independent identical distributions, is usually all that we need to capture exchangeability in practice

Of course, we can come up with a simple counter example to this:

Example I.32

Consider the probabilities of a given die landing on each of its six faces. The probabilities $\theta_1, \ldots, \theta_6$ are exchangeable, but the six parameters are constrained to sum to 1 and so cannot be modeled with a mixture of independent identical distributions.

I.D Hyper Priors and Hierarchical Models

Often, observations are not fully exchangeable, but partially or conditionally exchangeable

Partial Exchangeability

If observations can be grouped, we may make a hierarchical model, where each group has its own submodel, but the group properties are unknown. Assuming the group properties are exchangeable, we can use a common prior for them.

Conditional Exchangeability

If y_i has additional information x_i so that y_i are not exchangeable but (y_i, x_i) still are exchangeable, then we can make a joint model for (y_i, x_i) or a conditional model for $y_i|x_i$

I.D Hyper Priors and Hierarchical Models

In general, the usual way to model exchangeability with covariates is through conditional independence:

$$p(\theta_1, \ldots, \theta_J | x_1, \ldots, x_J) = \int \left[\prod_{j=1}^{J} p(\theta_j | \phi, x_j) \right] p(\phi | \mathbf{x}) d\boldsymbol{\phi}$$

In this way, exchangeable models become almost universally applicable because any information available to distinguish different units should be encoded in the x and y variables.

I.D Hyper Priors and Hierarchical Models

Example I.33

Let's consider a multi-center study of a new ointment for curing skin infections. The study took place over eight different centers. As the results may be center-specific, we may want to build a hierarchical model for this data.

Let n_j and θ_j be the number of subjects at center j and the probability of success at center j (where success is defined as the ointment cured the infection)

The number of cured infections at each center can be modeled by

$$y_j \sim Bin(n_j, \theta_j)$$

I.D Hyper Priors and Hierarchical Models

Despite the model having different parameters for each center, we may assume that each θ_j are iid samples from a beta distribution:

$$\theta_j \sim Beta(\alpha, \beta)$$

We then want to place a noninformative hyperprior on α and β

For now, let's just consider the likelihood and prior for this model

I.D Hyper Priors and Hierarchical Models

To recap, our likelihood is then

$$\mathcal{L}(y|\theta, \alpha, \beta) \propto \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j - y_j}$$

with conjugate prior on θ

$$\pi(\theta|\alpha, \beta) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}$$

and a noninformative hyperprior has the form

$$\pi(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$$

I.D Hyper Priors and Hierarchical Models

We now consider the normal model with exchangeable parameters

Example I.34

Suppose we have J independent experiments with experiment j estimating the parameter θ_j from n_j independent normally distributed data points, y_{ij}, each with known variance σ^2, thus

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2)$$

for $i = 1, \dots, n_j$ and $j = 1, \dots, J$.

I.D Hyper Priors and Hierarchical Models

Define the sample mean and sample variance for each group as:

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

for the sample mean for group j and

$$\sigma_j^2 = \sigma^2/n_j$$

for the sample variance for group j

Note that the $\bar{y}_j$ are sufficient statistics for the means θ_j

I.D Hyper Priors and Hierarchical Models

For the convenience of conjugacy, we assume the parameters θ_j come from a normal distribution with hyperparameters μ and τ giving the conjugate prior

$$\pi(\theta_1, \ldots, \theta_J | \mu, \tau) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\tau}} \exp\left[-\frac{1}{2\tau^2}(\theta_j - \mu)^2\right]$$

That is, we assume the θ_j's are conditionally independent given μ and τ

I.D Hyper Priors and Hierarchical Models

We then assign the noninformative hyperprior

$$\pi(\mu, \tau) \propto \pi(\tau)$$

In other words, we essentially assume that the conditional hyperprior on $\mu|\tau$ is a constant:

$$\pi(\mu, \tau) = \pi(\mu|\tau)\pi(\tau) \propto \pi(\tau)$$

Gelman notes that being vague in this context is OK since the combined data from the J experiments *should* be highly informative

I.D Hyper Priors and Hierarchical Models

Combining, we get the joint posterior:

$$
\begin{aligned}
p(\theta, \mu, \tau|y) &\propto \mathcal{L}(y|\theta)\pi(\theta|\mu, \tau)\pi(\mu, \tau) \\
&\propto \pi(\tau) \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\tau}} \exp\left[-\frac{1}{2\tau^2}(\theta_j - \mu)^2\right] \\
&\times \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left[-\frac{1}{2\sigma_j^2}(\bar{y}_j - \theta_j)^2\right]
\end{aligned}
$$

Conditional on the hyperparameters, and via conjugacy, we have the product of J independent unknown normal means

I.D Hyper Priors and Hierarchical Models

On other words, the conditional posterior distributions of the θ_j's, given μ and τ are

$$\theta_j | \mu, \tau, y \sim N(\hat{\theta}_j, \Sigma_j)$$

where

$$\hat{\theta}_J = \frac{\frac{1}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \text{ and } \Sigma_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

I.D Hyper Priors and Hierarchical Models

On other words, the conditional posterior distributions of the θ_j's, given μ and τ are

$$\theta_j|\mu,\tau,y \sim N(\hat{\theta}_j, \Sigma_j)$$

where

$$\hat{\theta}_J = \frac{\frac{1}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \text{ and } \Sigma_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

Question: how do we estimate μ and τ?

I.D Hyper Priors and Hierarchical Models

Consider the marginal posterior of the hyperparameters:

$$p(\mu, \tau) \propto \pi(\mu, \tau) p(y|\mu, \tau)$$

where the last term amounts to a ‘marginal likelihood’ factor

In many models, we cannot determine the ‘marginal likelihood’ factor so this representation is not often of use

However in the normal setting, $p(y|\mu, \tau)$ has a simple form

I.D Hyper Priors and Hierarchical Models

The marginal distributions of the group means, $\bar{y}_j$, averaging over θ, are independent (but not id) normals:

$$\bar{y}_j|\mu, \tau \sim N(\mu, \sigma_j^2 + \tau^2)$$

Thus the joint marginal posterior density is

$$p(\mu, \tau|y) \propto \pi(\mu, \tau) \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi(\sigma_j^2 + \tau^2)}} \exp\left[-\frac{1}{2(\sigma_j^2 + \tau^2)}(\bar{y}_j - \mu)^2\right]$$

Of course this form isn't the most useful, we can go further

I.D Hyper Priors and Hierarchical Models

Posterior Conditional of$\mu|\tau$

Conditional on τ and using the uniform prior density on $\pi(\mu|\tau)$, we get

$$\mu|\tau, y \sim N(\hat{\mu}, \Sigma_\mu)$$

where

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{y}_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \text{ and } \Sigma_\mu = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}\right)^{-1}$$

I.D Hyper Priors and Hierarchical Models

Marginal Posterior of τ

We can now obtain the posterior of τ analytically using conditional probability:

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \\ &\propto \frac{\pi(\tau)\prod_j \frac{1}{\sqrt{2\pi(\sigma_j^2+\tau^2)}}\exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_j-\mu)^2\right]}{\prod_j \frac{1}{\sqrt{2\pi\Sigma_\mu}}\exp\left[-\frac{1}{2\Sigma_\mu}(\mu-\hat{\mu})^2\right]} \end{aligned}$$

I.D Hyper Priors and Hierarchical Models

If we set μ equal to $\hat{\mu}$, we can simplify the expression

$$\begin{aligned} p(\tau|y) &\propto \frac{\pi(\tau)\prod_j \frac{1}{\sqrt{2\pi(\sigma_j^2+\tau^2)}} \exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_j-\mu)^2\right]}{\prod_j \frac{1}{\sqrt{2\pi\Sigma_\mu}} \exp\left[-\frac{1}{2\Sigma_\mu}(\hat{\mu}-\hat{\mu})^2\right]} \\ &\propto \pi(\tau)\Sigma_\mu^{1/2}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left[-\frac{(\bar{y}_j-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right] \end{aligned}$$

where μ and Σ_μ are as previously defined

I.D Hyper Priors and Hierarchical Models

Note that we have yet to specify a hyperprior for τ

Hyperprior for τ

If we wish to use a noninformative prior, we could simply select

$$\pi(\tau) \propto 1$$

This should result in a proper posterior

Alternatively, we could select a prior based on a scaled inverse-χ^2 distribution which could allow for an informative or weakly informative prior

I.D Hyper Priors and Hierarchical Models

Computation for Hierarchical Models

Provided we can identify the posterior conditional distributions of θ and the posterior marginal distribution of $\boldsymbol{\phi}$ is something we can draw from, we can take samples by first drawing $\boldsymbol{\phi}$ and then using it to draw θ

If anything is intractable, then we are at an impasse

Our next step will be to consider techniques for sampling from models with unrecognizable posteriors