STAT 532: Bayesian Data Analysis

Class 1: August 29, 2016

Today:

- Class Introductions
- Course overview
- Class Survey (Quiz 1)

Class 2: August 31, 2016

Last Time:

- Syllabus Day
- Class Survey (Quiz 1)

Today:

- Statistical Thought Experiments
- Bayesian mechanics

For Next Time:

- Read: Gelman's philosophy paper: Philosophy and the practice of Bayesian statistics

---

Experiment. Olympics Testing
Assume you were hired by the World Anti-Doping Agency to test olympic athletes for performance enhancing drugs. You are given results from eleven olympic champions, all of which test negative. Would do you believe the true probability of olympic champions using performance enhancing drugs would be?
Most likely non-zero, but as a fan I'd hope it would be small.

Calculate the maximum likelihood estimate of p the probability of Olympic champions using performance enhancing drugs.

Let y_i be 1 if test i is positive and zero otherwise, then $y = \sum_i y_i$.

$$\begin{aligned} \mathcal{L}(p|y) &\propto p^y(1-p)^{n-y} \\ \log \mathcal{L}(p|y) &\propto y\log(p) + (n-y)\log(1-p) \\ \frac{\delta \log \mathcal{L}(p|y)}{\delta p} &\propto y\log p + (n-y)\log(1-p) \text{ set} = 0 \\ \hat{p}_{MLE} &= \frac{y}{n} \end{aligned}$$

Is there a disconnect between your earlier belief and the estimated probability using MLE?

Now we will talk about the mechanics of Bayesian statistics and revisit the olympic testing problem.
Sampling Model: The sampling model $p(y|\theta)$, where θ is a parameter describes the belief that y would be the outcome of the study, given θ was known.
Ex. Binomial Model. $p(y|p) = \binom{n}{y} p^y(1-p)^{n-y}$

Likelihood Function: The likelihood function $\mathcal{L}(\theta|y)$ is proportional to the sampling model, but is a function of the parameters. When using a likelihood function, typically the normalizing constants are dropped.
Ex. Binomial Model. $\mathcal{L}(p|y) \propto p^y(1-p)^{n-y}$

Prior Distribution: The prior distribution $p(\theta)$ describes the degree of belief over the parameter space Θ.
Ex. Beta Distribution. $p(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$ One example is Beta(1,1), Uniform Model. $p(p) = 1$, $p \in [0,1]$, $p = 0$ otherwise. Note $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Posterior Distribution: Given a prior distribution and a likelihood function, or sampling model, the posterior distribution of the parameters can be calculated using Bayes' rule.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})} \tag{1}$$

In Bayesian statistics, inferences are made from the posterior distribution. In cases where analytical solutions are possible, the entire posterior distribution provides an informative description of the uncertainty present in the estimation. In other cases credible intervals are used to summarize the uncertainty in the estimation.

Experiment. Olympic Testing (with Bayes).
Now reconsider the olympic testing program from a Bayesian perspective. Use the Beta(α,β) as the prior

distribution for p and compute the posterior distribution for p.

$$p(p|y) = \frac{p(y|p)p(p)}{\int p(y|p)p(p)dp} = \frac{\frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{\int \frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}dp}$$

first the integration in the denominator.

$$\begin{aligned}
\int \frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}dp &= \frac{\binom{n}{y}}{B(\alpha,\beta)}\int p^{\alpha+y-1} + (1-p)^{\beta+n-y-1)}dp \\
&= \frac{\binom{n}{y}B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}\int \frac{p^{\alpha+y-1} + (1-p)^{\beta+n-y-1}}{B(\alpha+y,\beta+n-y}dp \\
&= \frac{\binom{n}{y}B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}
\end{aligned}$$

Now the posterior distribution follows as:

$$\begin{aligned}
p(p|y) &= \frac{\frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{\frac{\binom{n}{y}B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}} \\
&= \frac{p^{\alpha+y-1} + (1-p)^{\beta+n-y-1}}{B(\alpha+y,\beta+n-y)} \\
&\sim Beta(\alpha+y,\beta+n-y).
\end{aligned}$$

Now use a Beta(1,1) distribution as the prior for p(p) and compute $p(p|y)$. Then $p(p|y) \sim Beta(1,14)$, where $E[p(p|y)] = 1/14$.
How do these results compare with your intuition which we stated earlier?

How about the MLE estimate?

Class 3: September 2, 2016

Last Time:

- Olympic testing problem
- Bayesian mechanics

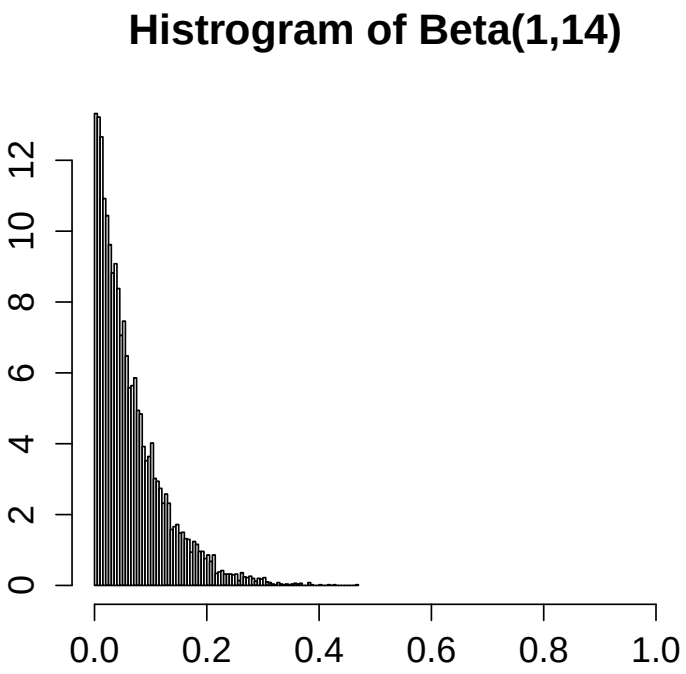


Figure 1: Here is a histogram representation of the posterior distribution.

Assigned Reading

- Gelman's philosophy paper: Philosophy and the practice of Bayesian statistics

Today:

- Discussion of Gelman's paper
- a *brief* overview of philosophy of Bayesian statistics

Classical, or frequentist, statistical paradigm:

- Estimate fixed parameters by maximizing the likelihood function $\mathcal{L}(\theta|X)$.
- Uncertainty in estimates is expressed using **confidence.** The concept of confidence requires a frequentist viewpoint. Specifically, a confidence interval states that if an experiment was conducted a large number of times, we'd expect the true estimate to be contained in the interval at the specificed level. No probabalistic statements can be made (e.g. the probability the true value is in the specified confidence interval).
- Inference is often conducted using hypothesis testing and requires requires **p-values**. Conceptually, a p-value is the probabilty of obtaining the result (or a more extreme outcome) given the stated hypothesis. However, in recent years, the use of p-values has been under increased scrutiny. We will dedicate a class to this later in the semester.

Bayesian statistical paradigm

- Given a stated prior $p(\theta)$ a posterior distribution is computed for the parameters $p(\theta|X)$.
- Uncertainty in the estimates is expressed using the posterior distribution. Often the posterior is summarized by a **credible interval** which does allow probabilistic statements (e.g. the probability the true value is contained in the credible interval.)
- For inference, the posterior distribution is typically summarized using a credible interval and often combined with *maximum a posteriori* (MAP) point estimates.

Class 4: September 7, 2016

Last Time:

- Philosophy of Bayesian Statistics

Today:

- Start: Hoff - Chapter 2. Probability Theory in two days.

Axioms of Probability

Kolmogorov's Axioms of Probability:

1. $0 = Pr(\text{not } H|H) \leq Pr(F|H) \leq Pr(H|H) = 1$
2. $Pr(F \cup G|H) = Pr(F|H) + Pr(G|H)$ if $F \cap G = \emptyset$
3. $Pr(F \cap G|H) = Pr(G|H)Pr(F|G \cap H)$

The textbook discusses the idea of belief functions. The important thing to note here is that a probability function can be used to express beliefs in a principled manner.

Partitions and Bayes Rule

A collection of sets $\{H_1, ..., H_K\}$ is a **partition** of another set $\mathcal{H}$ if

1. the events are disjoint, $H_i \cap H_j = \emptyset$ for $i \neq j$ and
2. the union of the sets is $\mathcal{H}$, $\cup_{k=1}^{K} H_k = \mathcal{H}$.

Examples

- Let $\mathcal{H}$ be a Bozemanite's favorite ski hill. Partitions include:
  - {Big Sky, Bridger Bowl, other}
  - {Any Montana Ski Hill, Any Colorado Ski Hill, other}
- Let $\mathcal{H}$ be the number of stats courses taken by people in this room. Partitions include:
  - {0, 1, 2, ...}

– {0 - 3, 4-6, 7-10, 10+}

Suppose $\{H_1, ..., H_K\}$ is a partition of $\mathcal{H}$, $Pr(\mathcal{H}) = 1$, and E is some specific event. The axioms of probability imply the following statements:

1. **Rule of Total Probability:** $\sum_{k=1}^{K} Pr(H_k) = 1$

2. **Rule of Marginal Probability:**

$$\begin{aligned} Pr(E) &= \sum_{k=1}^{K} Pr(E \cap H_k) \\ &= \sum_{k=1}^{K} Pr(E|H_k)Pr(H_k) \end{aligned}$$

3. **Bayes rule:**

$$\begin{aligned} Pr(H_j|E) &= \frac{Pr(E|H_j)Pr(H_j)}{Pr(E)} \\ &= \frac{Pr(E|H_j)Pr(H_j)}{\sum_{k=1}^{K} Pr(E|H_k)Pr(H_k)} \end{aligned}$$

Example. Assume a sample of MSU students are polled on their skiing behavior. Let $\{H_1, H_2, H_3, H_4\}$ be the events that a randomly selected student in this sample is in, the first quartile, second quartile, third quartile and 4th quartile in terms of number of hours spent skiing.
Then $\{Pr(H_1), Pr(H_2), Pr(H_3), Pr(H_4)\} = \{.25, .25, .25, .25\}$.
Let E be the event that a person has a GPA greater than 3.0.
Then $\{Pr(E|H_1), Pr(E|H_2), Pr(E|H_3), Pr(E|H_4)\} = \{.40, .71, .55, .07\}$.
Now compute the probability that a student with a GPA greater than 3.0 falls in each quartile for hours spent skiing : $\{Pr(H_1|E), Pr(H_2|E), Pr(H_3|E), Pr(H_4|E)\}$

$$\begin{aligned} Pr(H_1|E) &= \frac{Pr(E|H_1)Pr(H_1)}{\sum_{k=1}^{4} Pr(E|H_k)Pr(H_k)} \\ &= \frac{Pr(E|H_1)}{Pr(E|H_1) + Pr(E|H_2) + Pr(E|H_3) + Pr(E|H_4)} \\ &= \frac{.40}{.40 + .71 + .55 + .07} = \frac{.4}{1.73} = .23 \end{aligned}$$

Similarly, $Pr(H_2|E) = .41$, $Pr(H_3|E) = .32$, and $Pr(H_4|E) = .04$.

Independence

Two events F and G are conditionally independent given H if $Pr(F \cap G|H) = Pr(F|H)Pr(G|H)$.
If F and G are conditionally independent given H then:

$$\begin{aligned} Pr(G|H)Pr(F|H \cap G) &= \text{(always true) } Pr(F \cap G|H) = \text{(under indep)} Pr(F|H)Pr(G|H) \\ Pr(G|H)Pr(F|H \cap G) &= Pr(F|H)Pr(G|H) \\ Pr(F|H \cap G) &= Pr(F|H) \end{aligned}$$

Example. What is the relationship between $Pr(F|H)$ and $Pr(F|H \cap G)$ in the following situation?
$F =$\{ you draw the jack of hearts \}
$G =$\{ a mind reader claims you drew the jack of hearts \}
$H =$\{ the mind reader has extrasensory perception \}
$Pr(F|H) = 1/52$ and $Pr(F|G \cap H) > 1/52$.

Class 5: September 9, 2016

Last Time:

- Hoff - Chapter 2. Bayes Rule, Conditional Independence

Today:

- More: Hoff - Chapter 2. Random Variables, Distributions

Assignments:

- HW 1, due Sept. 16

Random Variables

In Bayesian inference a random variable is defined as an unknown numerical quantity about which we make probability statements. For example, the quantitative outcome of a study is performed. Additionally, a fixed but unknown population parameter is also a random variable.

Discrete Random Variables

Let Y be a random variable and let $\mathcal{Y}$ be the set of all possible values of Y. Y is discrete if the set of possible outcomes is countable, meaning that $\mathcal{Y}$ can be expressed as $\mathcal{Y} = \{y_1, y_2, ...\}$.

The event that the outcome Y of our study has the value y is expressed as $\{Y = y\}$. For each $y \in \mathcal{Y}$, our shorthand notation for $Pr(Y = y)$ will be $p(y)$. This function (known as the probability distribution function (pdf)) $p(y)$ has the following properties.

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$

Example 1. Binomial Distribution
Let $\mathcal{Y} = \{0, 1, 2, ..n\}$ for some positive integer n. Then $Y \in \mathcal{Y}$ has a binomial distribution with probability θ if

$$Pr(Y = y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}.$$

Example 2. Poisson Distribution
Let $\mathcal{Y} = \{0, 1, 2...\}$. Then $Y \in \mathcal{Y}$ has a Poisson distribution with mean θ if

$$Pr(Y = y|\theta) = \theta^y \exp(-\theta)/y!$$

Continuous Random Variables

Suppose that the sample space $\mathcal{Y}$ is $\mathbb{R}$, then $Pr(Y \leq 5) \neq \sum_{y \leq 5} p(y)$ as this sum does not make sense. Rather define the cumulative distribution function (cdf) $F(y) = Pr(Y \leq y)$. The cdf has the following properties:

1. $F(\infty) = 1$
2. $F(-\infty) = 0$
3. $F(b) \leq F(a)$ if $b < a$.

Probabilities of events can be derived as:

- $Pr(Y > a) = 1 - F(a)$
- $Pr(a < Y < b) = F(b) - F(a)$

If F is continuous, then Y is a continuous random variable. Then $F(a) = \int_{-\infty}^{a} p(y)dy$.
Example. Normal distribution.
Let $\mathcal{Y} = (-\infty, \infty)$ with mean μ and variance σ^2. Then y follows a normal distribution if

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

Moments of Distributions

The mean or expectation of an unknown quantity Y is given by

$$\begin{aligned} E[Y] &= \sum_{y\in\mathcal{Y}} yp(y) \text{ if Y is discrete and} \\ E[Y] &= \int_{y\in\mathcal{Y}} yp(y) \text{ if Y is continuous.} \end{aligned}$$

The variance is a measure of the spread of the distribution.

$$\begin{aligned} Var[Y] &= E[(Y - E[Y])^2] \\ &= E[Y^2 - 2YE[Y] + E[Y]^2] \\ &= E[Y^2] - 2E[Y]^2 + E[Y]^2 \\ &= E[Y^2] - E[Y]^2 \end{aligned}$$

If $Y \sim$ Binomial(n, p), then $E[Y] = np$ and $Var[Y] = np(1-p)$.
if $Y \sim$ Poisson(μ), then $E[Y] = \mu$ and $Var[Y] = \mu$.
if $Y \sim$ Normal(μ, σ^2), then $E[Y] = \mu$ and $Var[Y] = \sigma^2$.

Joint Distributions

Let Y_1, Y_2 be random variables, then the joint pdf or joint density can be written as

$$P_{Y_1,Y_2}(y_1, y_2) = Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \text{ for } y_1 \in \mathcal{Y}_1, y_2 \in \mathcal{Y}_2$$

The marginal density of Y_1 can be computed from the joint density:

$$\begin{aligned} p_{Y_1}(y_1) &= Pr(Y_1 = y_1) \\ &= \sum_{y2\in\mathcal{Y}_2} Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \\ &= \sum_{y2\in\mathcal{Y}_2} p_{Y_1,Y_2}(y_1, y_2). \end{aligned}$$

Note this is for discrete random variables, but a similar derivation holds for continuous.

The conditional density of Y_2 given $\{Y_1 = y_1\}$ can be computed from the joint density and the marginal density.

$$\begin{aligned} p_{Y_2|Y_1}(y_2|y_1) &= \frac{Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{Pr(Y_1 = y_1)} \\ &= \frac{p_{Y_1,Y_2}(y_1, y_2}{p_{Y_1}(y_1)} \end{aligned}$$

Note the subscripts are often dropped, so $p_{Y_1,Y_2}(y_1, y_2) = p(y_1, y_2)$, ect...

Independent Random Variables and Exchangeability

Suppose $Y_1, ..., Y_n$ are random variables and that θ is a parameter corresponding to the generation of the random variables. Then $Y_1, ..., Y_n$ are conditionally independent given θ if

$$Pr(Y_1 \in A_1, ..., Y_n \in A_n|\theta) = Pr(Y_1 \in A_1) \times ... \times Pr(Y_n \in A_n)$$

where $\{A_1, ..., A_n\}$ are sets. Then the joint distribution can be factored as

$$p(y_1, ..., y_n|\theta) = p_{Y_1}(y1|\theta) \times ... \times p_{Y_n}(y_n|\theta).$$

If the random variables come from the same distribution then they are conditionally independent and identically distributed, which is noted $Y_1, ..., Y_n|\theta \sim i.i.d.p(y|\theta$ and

$$p(y_1, ..., y_n|\theta) = p_{Y_1}(y1|\theta) \times ... \times p_{Y_n}(y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

Exchangeability

Let $p(y_1, ...y_n)$ be the joint density of $Y_1, ..., Y_n$. If $p(y_1, ..., y_n) = p(y_{\pi_1}, ..., y_{\pi_n})$ for all permutations π of $\{1, 2, ...,n\}$, then $Y_1, ..., Y_n$ are exchangeable.

Example. Assume data has been collected on apartment vacancies in Bozeman. Let $y_i = 1$ if an *affordable* room is available. Do we expect $p(y_1 = 0, y_2 = 0, y_3 = 0, y_4 = 1) = p(y_1 = 1, y_2 = 0, y_3 = 0, y_4 = 0)$? If so the data are exchangeable.

Let $\theta \sim p(\theta)$ and if $Y_1, ..., Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ) $Y_1, ..., Y_n$ are exchangeable.
Proof omitted, see textbook for details.

Class 6: September 12, 2016

Last Time:

- Hoff - Chapter 2. Random Variables, Distributions
- HW 1 assigned

Today:

- One-parameter Models: Binomial Models (Ch. 3.1)

The binomial model

Example. After suspicious performance in the weekly soccer match, 37 mathematical sciences students, staff, and faculty were tested for performance enhancing drugs. Let $Y_i = 1$ if athlete i tests positive and $Y_i = 0$ otherwise. A total of 13 athletes tested positive.
Write the sampling model $p(y_1, ..., y_{37}|\theta)$.

$p(y_1, ..., y_{37}|\theta) = \binom{N}{y}\theta^y(1-\theta)^{n-y}$

Assume a uniform prior distribution on $p(\theta)$. Write the pdf for this distribution.

$$p(\theta) = \begin{cases} 1, & \text{if } 0 \leq \theta \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In what larger class of distributions does this distribution reside? What are the parameters?

Beta, $\alpha = 1$, $\beta = 1$.
Beta distribution. Recall, $\theta \sim Beta(\alpha, \beta)$ if:

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

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

Now compute the posterior distribution, $p(\theta|\boldsymbol{y})$.

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \frac{\mathcal{L}(\theta|\boldsymbol{y})p(\theta)}{\int_\theta \mathcal{L}(\theta|\boldsymbol{y})p(\theta)d\theta} \\ &= \frac{\mathcal{L}(\theta|\boldsymbol{y})p(\theta)}{p(\boldsymbol{y})} \\ &\propto \mathcal{L}(\theta|\boldsymbol{y})p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{13+1-1}(1-\theta)^{37-13+1-1} \\ &\sim Beta(14,25) \end{aligned}$$

The posterior expectation,$E[\theta|y] = \frac{\alpha+y}{\alpha+\beta+n}$, is a function of prior information and the data.

Conjugate Priors

We have shown that a beta prior distribution and a binomial sampling model lead to a beta posterior distribution. This class of beta priors is **conjugate** for the binomial sampling model.

Def: Conjugate *A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(y|\theta)$ if $p(\theta) \in \mathcal{P} \rightarrow p(\theta|y) \in \mathcal{P}$.*

Conjugate priors make posterior calculations simple, but might not always be the best representation of prior beliefs.

Predictive Distributions

An important element in Bayesian statistics is the predictive distribution, in this case let Y^* be the outcome of a future experiment. We are interested in computing:

$$\begin{aligned} Pr(Y^*=1|y_1,...,y_n) &= \int Pr(Y^*=1|\theta,y_1,...,y_n)p(\theta|y_1,...,y_n)d\theta \\ &= \int \theta p(\theta|y_1,...,y_n)d\theta \\ &= E[\theta|y_1,...,y_n] \\ &= \frac{\alpha+\boldsymbol{y}}{\alpha+\beta+n}, \text{where } \boldsymbol{y} = \sum_i^n y_i \end{aligned}$$

Note that the predictive distribution does not depend on any unknown quantities, but rather only the observed data. Furthermore, Y^* is not independent of $Y_1,...,Y_n$ but depends on them through θ.

Posterior Intervals

With a Bayesian framework we can compute **credible intervals**.
Credible Interval: *An interval* $[l(y), u(y)]$ *is an* $1-\alpha\%$ *credible interval for* θ *if:*

$$Pr(l(y) < \theta < u(y)|Y = y) = 1 - \alpha \tag{3}$$

Recall in a frequentist setting

$$Pr(l(y) < \theta < u(y)|\theta) = \begin{cases} 1, & \text{if } \theta \in [l(y), u(y)] \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Note that in some settings Bayesian intervals can also have frequentist coverage probabilities, at least asymptotically.

Quantile based intervals
With quantile based intervals, the posterior quantiles are used with $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ such that:

1. $Pr(\theta < \theta_{\alpha/2}|Y = y) = \alpha/2$ and
2. $Pr(\theta > \theta_{1-\alpha/2}|Y = y) = \alpha/2$.

Quantile based intervals are typically easy to compute.

Highest posterior density (HPD) region
A 100 $\times$ *(1-*α*)*$\%$ *HPD region consists of a subset of the parameter space,* $s(y) \subset \Theta$ *such that*

1. $Pr(\theta \in s(y)|Y = y) = 1 - \alpha$
2. *If* $\theta_a \in s(y)$*, and* $\theta_b \notin s(y)$*, then* $p(\theta_a|Y = y) > p(\theta_b|Y = y)$.

All points in the HPD region have higher posterior density than those not in region. Additionally the HPD *region* need not be a continuous interval. HPD regions are typically more computationally intensive to compute than quantile based intervals.

Class 7: September 14, 2016

Last Time:

- The binomial model

Today:

- One-parameter Models: Poisson Models (Ch. 3.2) and Exponential Family (Ch 3.3)

Next Time:

- HW 1 due

The Poisson Model

Recall, $Y \sim Poisson(\theta)$ if

$$Pr(Y = y|\theta) = \frac{\theta^y \exp(-\theta)}{y!}. \tag{5}$$

Properties of the Poisson distribution:

- $E[Y] = \theta$
- $Var(Y) = \theta$
- $\sum_i^n Y_i \sim Poisson(\theta_1 + ... + \theta_n)$ if $Y_i \sim Poisson(\theta_i)$

Conjugate Priors for Poisson

Recall conjugate priors for a sampling model have a posterior model from the same class as the prior. Let $y_i \sim Poisson(\theta)$, then

$$\begin{aligned} p(\theta|y_1, ..., y_n) &\propto p(\theta)\mathcal{L}(\theta|y_1, ..., y_n) &(6)\\ &\propto p(\theta) \times \theta^{\sum_{y_i}} \exp(-n\theta) &(7)\end{aligned}$$

Thus the conjugate prior class will have the form $\theta^{c_1} \exp(c_2\theta)$. This is the kernel of a gamma distribution.

A positive quantity θ has a $Gamma(a, b)$ distribution if:

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1} \exp(-b\theta), \text{ for } \theta, a, b > 0$$

Properties of a gamma distribution:

- $E[\theta] = a/b$
- $Var(\theta) = a/b^2$

Posterior Distribution

Let $Y_1, ..., Y_n \sim Poisson(\theta)$ and $p(\theta) \sim gamma(a, b)$, then

$$\begin{aligned} p(\theta|y_1, ..., y_n) &= \frac{p(\theta)p(y_1, ..., y_n|\theta)}{p(y_1, ..., y_n)} && (8) \\ &= \{\theta^{a-1}\exp(-b\theta)\} \times \{\theta^{\sum y_i}\exp(-n\theta)\} \times \{c(y_1, ..., y_n, a, b)\} && (9) \\ &\propto \theta^{a+\sum y_i - 1}\exp(-\theta(b+n)) && (10) \\ & && (11) \end{aligned}$$

So $\theta|y_1, ..., y_n \sim gamma(a + \sum y_i, b + n)$. Note that

$$\begin{aligned} E[\theta|y_1, ..., y_n] &= \frac{a + \sum y_i}{b+n} \\ &= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n} \end{aligned}$$

So now a bit of intuition about the prior distribution. The posterior expectation of θ is a combination of the prior expectation and the sample average:

- b in interpreted as the number of prior observations
- a is interpreted as the sum of the counts from b prior observations

When $n >> b$ the information from the data dominates the prior distribution.

Predictive distribution

The predictive distribution, $p(y^*|y_1, ..., y_n)$, can be computed as:

$$\begin{aligned} p(y^*|y_1, ..., y_n) &= \int p(y^*|\theta, y_1, ..., y_n)p(\theta|y_1, ..., y_n)d\theta \\ &= \int p(y^*|\theta)p(\theta|y_1, ..., y_n)d\theta \\ &= \int \left\{\frac{\theta^{y^*}\exp(-\theta)}{y^*!}\right\}\left\{\frac{(b+n)^{a+\sum y_i}}{\Gamma(a+\sum y_i)}\theta^{a+\sum y_i - 1}\exp(-(b+n)\theta)\right\} \\ &= ... \\ &= ... \end{aligned}$$

You can (and likely will) show that $p(y^*|y_1, ..., y_n) \sim NegBinom(a + \sum y_i, b + n)$.

Exponential Families

The binomial and Poisson models are examples of one-parameter exponential families. A distribution follows a one-parameter exponential family if it can be factorized as:

$$p(y|\theta) = h(y)c(\phi)\exp(\phi t(y)), \tag{12}$$

where ϕ is the unknown parameter and $t(y)$ is the sufficient statistic. The using the class of priors, where $p(\phi) \propto c(\phi)^{n_0}\exp(n_0 t_0 \phi)$, is a conjugate prior. There are similar considerations to the Poisson case where n_0 can be thought of as a "prior sample size" and t_0 is a "prior guess."

Class 8: September 16, 2016

Last Time:

- The Poisson model and exponential families

Today:

- HW 1 due
- Ideas about prior distributions
- HW 2 assigned

Noninformative Priors

A **noninformative prior**, $p(\theta)$, contains no information about θ.

Example 1. Suppose θ is the probability of success in a binomial distribution, then the uniform distribution on the interval $[0, 1]$ is a noninformative prior.

Example 2. Suppose θ is the mean parameter of a normal distribution. What is a noninformative prior distribution for the mean?

- One option would be a Normal distribution centered at zero with very large variance. However, this will still contain more mass at the center of the distribution and hence, favor that part of the parameter space.
- We'd like to place a uniform prior on θ, but $\int_{\infty}^{\infty} \text{c}\ dx = \infty$, so the uniform prior on the real line is not a probability distribution. Does this matter? This was actually a common prior used by LaPlace.

  Sometimes is the answer. Ultimately the inference is based on the posterior, so if an improper prior leads to a proper posterior that is okay. In most simple analyses we will see in this class improper priors will be fine.

Invariant Priors

Recall ideas of variable transformation (from Casella and Berger): *Let X have pdf $p_x(x)$ and let Y = g(X), where g is a monotone function. Suppose $p_X(x)$ is continuous and that $g^{-1}(y)$ has a continuous derivative. Then the pdf of Y is given by*

$$p_y(y) = p_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

Example. Let $p_x(x) = 1$, for $x \in [0, 1]$ and let $Y = g(x) = -\log(x)$, then

$$\begin{aligned} p_y(y) &= p_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \exp(-y) \text{ for } y \in [0, \infty) \end{aligned}$$

Now if $p_x(x)$ had been a prior on X, the transformation to y and $p_y(y)$ results in an informative prior for y.

Jeffreys Priors

The idea of invariant priors was addressed by Jeffreys. Let $p_J(\theta)$ be a Jeffreys prior if:

$$p_J(\theta) = [I(\theta)]^{1/2},$$

where $I(\theta)$ is the expected Fisher information given by

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

Example. Consider the Normal distribution and place a prior on μ when σ^2 is known. Then the Fisher information is

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2}{\partial \mu^2}\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)\right] \\ &= \frac{1}{\sigma^2} \end{aligned}$$

Hence in this case the Jeffreys prior for μ is a constant. A similar derivation for the joint prior $p(\mu, \sigma) = \frac{1}{\sigma}$

Advantages and Disadvantages of Objective Priors

Advantages

- Objective prior distributions reflect the idea of there being very little information available about the underlying process.
- There are sometimes mathematically equivalent results obtained by Bayesian methods using objective prior distributions and results obtained using frequentist methods on the same problem (but the results in the two cases have different philosophical interpretations.)
- Objective prior distributions are easy to defend.

Disadvantages

- Sometimes improper priors result from objective prior distributions
- In multiple parameter situations, the parameters are often taken to be independent
- Improper objective prior distributions are problematic in Bayes factor computations and some model selection settings.

Advantages and Disadvantages of Subjective Priors

Advantages

- Subjective prior distributions are proper.

- Subjective prior distributions introduced informed understanding into the analysis.
- Subjective prior distributions can provide sufficient information to solve problems when other methods are not sufficient.

Disadvantages

- It is difficult to assess subjective prior beliefs as it is hard to translate prior knowledge into a probability distribution
- Result of a Bayesian analysis may not be relevant if the prior beliefs used do not match your own.
- A subjective prior may not be computationally convenient

In many cases weakly-informative prior distributions are used that have some of the benefits of subjective priors without imparting strong information into the analysis.

Class 9: September 19, 2016

Last Time:

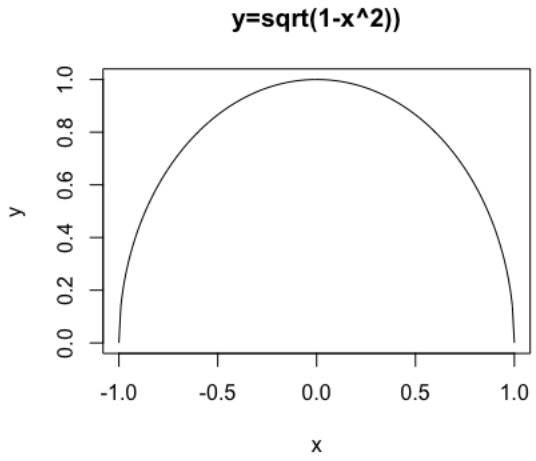
- Choosing priors

This week:

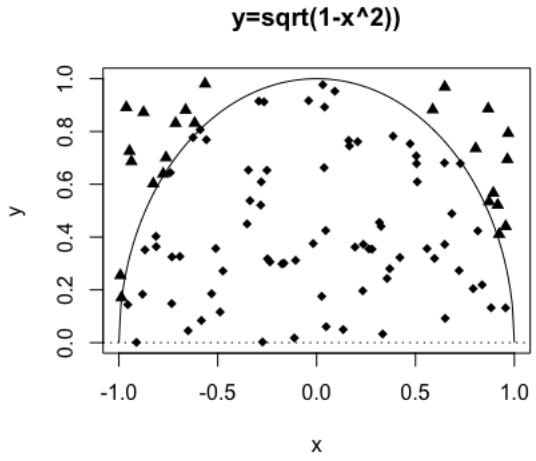
Overview of Computational Statistics

- Monte Carlo overview
- Posterior inference
- Sampling from predictive distributions
- Posterior predictive model checking

Exercise 1. Consider the function $g(x) = \sqrt{1-x^2}$, where $x \in [0,1]$. The goal is to estimate $I = \int g(x)dx$.



One way to do this is to simulate points from a uniform distribution with a known area. Then we compute the proportion of points that fall under the curve for $g(x)$. Specifically, we create samples from a uniform function on $[-1, 1]$. The area under this function is 2. To compute I we estimated the proportion of responses in I that are also under $g(x)$.




```
> x <- seq(-1,1, by=.01)
> y <- sqrt(1-x^2)
> plot(x,y,type='l', main= "y=sqrt(1-x^2))")
> abline(h=0,lty=3)
> num.sims <- 100000
> f.x <- runif(num.sims,-1,1)
> f.y <- runif(num.sims)
```

```
> accept.samples <- (1:num.sims)[f.y < sqrt(1-f.x^2)]
> #points(f.x[accept.samples],f.y[accept.samples],pch=18)
> reject.samples <- (1:num.sims)[f.y > sqrt(1-f.x^2)]
> #points(f.x[reject.samples],f.y[reject.samples],pch=17)
> accept.proportion <- length(accept.samples)/num.sims
> area <- 2 * accept.proportion; area
[1] 1.57014
```

This matches with the analytical solution of $\pi/2$.

Monte Carlo Procedures

Monte Carlo procedures use random sampling to estimate mathematical or statistical quantities. These computational algorithms are defined by running for a fixed time (number of samples/iterations) and result in a random estimate. There are three main uses for Monte Carlo procedures: 1.) optimization, 2.) integration, and 3.) generating samples from a probability distribution.

Monte Carlo methods were introduced by John von Neumann and Stanislaw Ulam at Los Alamos. The name Monte Carlo was a code name referring the Monte Carlo casino in Monaco. Monte Carlo methods were central to the Manhattan project and continued development of physics research related to the hydrogen bomb.

An essential part of many scientific problems is the computation of the integral, $I = \int_{\mathcal{D}} g(x)dx$ where $\mathcal{D}$ is often a region in a high-dimensional space and $g(x)$ is the target function of interest. If we can draw independent and identically distributed random samples $x_1, x_2, \ldots, x_n$ uniformly from $\mathcal{D}$ (by a computer) an approximation to I can be obtained as

$$\hat{I}_n = \frac{1}{n}\left(g(x_1) + \cdots + g(x_n)\right).$$

The law of large numbers states that the average of many independent random variables with common mean and finite variances tends to stabilize at their common mean; that is

$$\lim_{n \leftarrow \infty} \hat{I}_n = I, \text{ with probability 1.}$$

A related procedure that you may be familiar with is the *Riemann approximation*. Consider a case where $\mathcal{D} = [0, 1]$ and $I = \int_0^1 g(x)dx$ then

$$\tilde{I} = \frac{1}{n}\left(g(b_1) + \cdots + g(b_n)\right),$$

where $b_i = i/n$. Essentially this method is a grid based evaluation. This works will for a smooth function in low dimension, but quickly runs into the "curse of dimensionality".

Accept - Reject Sampling

In many scenarios, sampling from a function $g(x)$ can be challenging (in that we cannot use `rg.function()` to sample from it. The general idea of accept reject sampling is to simulation observations from another distribution $f(x)$ and accept the response if it falls under the distribution $g(x)$.

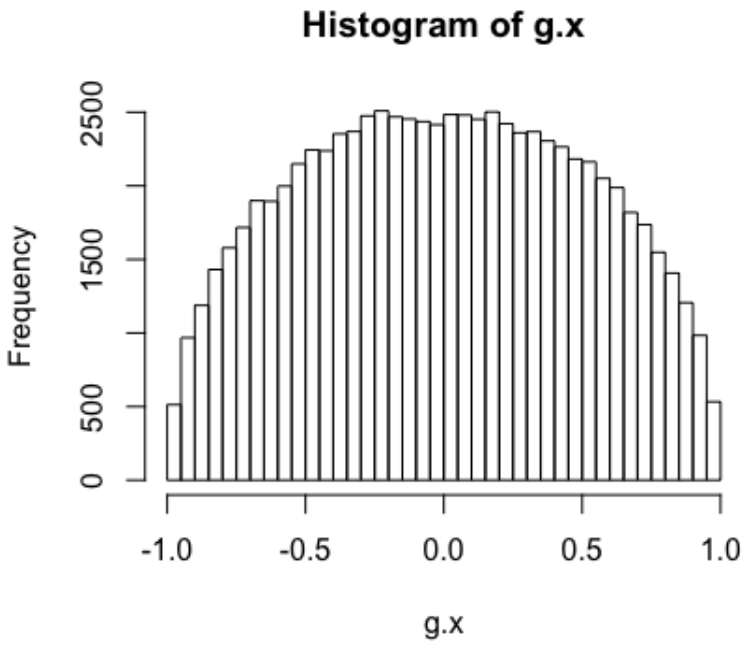
Formally the algorithm for the Accept-Reject Method follows as:

1. Generate $X \sim f, U \sim Unif[0,1]$
2. Accept $Y = X$ if $U \leq g(x)/Mf(x)$,

where $f(x)$ and $g(x)$ are normalized probability distributions and M is a constant ≥ 1.

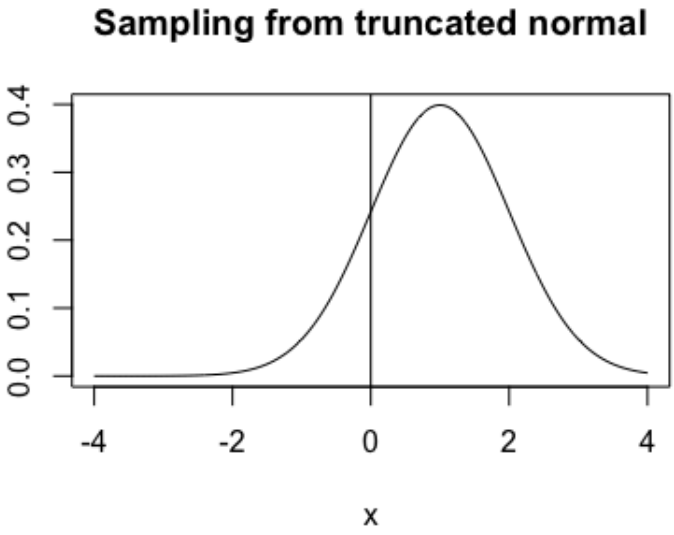
Similar to the context of the problem above, suppose we want to draw samples from a normalized form of $g(x) = \frac{\sqrt{1-x^2}}{\pi/2}$. Then $f(x) = \frac{1}{2}$ for $x \in [-1,1]$ and $M = \frac{4}{\pi}$.

```
x <- runif(num.sims,-1,1) # simulate samples
u.vals <- runif(num.sims)
M <- 4/pi
f.x <- 1/2

accept.ratio <- (sqrt(1-x^2) / (pi/2)) / (f.x*M) # g/fM
accept.index <- (1:num.sims)[u.vals < accept.ratio]
g.x = x[accept.index]
hist(g.x,breaks='FD')
```



Now we have samples from $g(x)$ and could, for example, compute $I = \int xg(x)dx$

Exercise 2. Without the use of packages such as `rtnorm()` how would you draw a sample from a truncated normal distribution?



Again we will use accept-reject sampling. Simulate points from a normal distribution with the same mean and variance. Let M = $\frac{1}{\Phi(c;\mu,\sigma^2)}$, where $\Phi(.;\mu,\sigma^2)$ is the cdf function a normal random variable and c is the truncation point. Then all values of x greater than c are accepted and all values less than c are rejected. We will use this formulation for binary Bayesian regression models using a probit link.

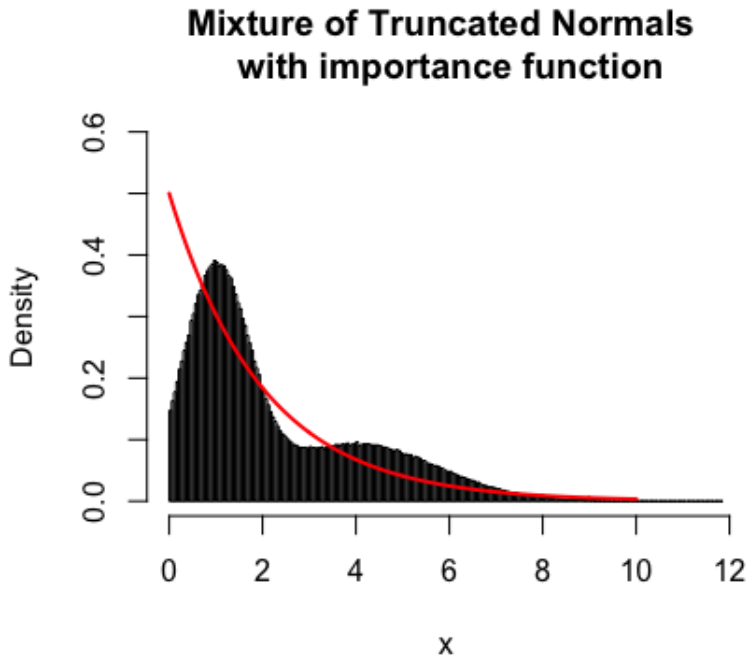
Importance Sampling

Importance sampling is related to the idea of accept-reject sampling, but emphasizes focusing on the "important" parts of the distribution and uses all of the simulations. In accept-reject sampling, computing the value M can be challenging in its own right. We can alleviate that problem with importance sampling.

Again the goal is to compute some integral, say $I = \int h(x)g(x)dx = E[h(x)]$. We cannot sample directly from $g(x)$ but we can evaluate the function. So the idea is to find a distribution that we can simulate observations from, $f(x)$, that ideally looks similar to $g(x)$. The importance sampling procedure follows as:

1. Draw $x_1, \ldots, x_n$ from trial distribution $f(x)$.
2. Calculate the *importance weight*:
   $w_j = g(x_j)/f(x_j)$, for $j = 1, \ldots, n$
3. Approximate $I = \frac{w_1 h(x_1) + \cdots + w_n h(x_n)}{w_1 + \cdots + w_n}$

Example. Compute the mean of a mixture of truncated normal distributions. Let $X \sim (.4)N(4,3)_{+(0)} + (.6)N(1,.5)_{+(0)}$.



Let the trial distribution be an Exponential(.5) distribution. Then the importance sampling follows as:

```
> w <- ( .6*dnorm(f,1,sqrt(.5))/ (1-pnorm(0,1,sqrt(.5))) +
+         .4*dnorm(f,4,sqrt(3))/ (1-pnorm(0,4,sqrt(3))) ) /
+        dexp(f,.5)
> sum(w*f) / sum(w)
[1] 2.286383
```

Note the `rtnorm()` function has this implementation using importance sampling with an exponential distribution.

Later in class we will spend a little time on Bayesian time series models using dynamic linear models. One way to fit these models is using a specific type of importance sampling in a sequential Monte Carlo framework, known as particle filters.

Monte Carlo Variation

To assess the variation and, more specifically, the convergence of Monte Carlo methods, the Central Limit Theorem is used. The nice thing about this setting is that samples are relatively cheap to come by, only requiring a little more computational run time.

A general prescription is to run the Monte Carlo procedure a few times and assess the variability between the outcomes. Increase the sample size, if needed.

Posterior Inference

Monte Carlo procedures fit naturally in the Bayesian paradigm for creating draws from the posterior distribution. A nice feature of Monte Carlo procedures and posterior inference is for assessing functions of the posterior.

Example. Consider a binomial model where we have a posterior distribution for the probability term, θ. To make inferences on the log-odds $\gamma = \log \frac{\theta}{1-\theta}$, the following procedure is used:

1. draw samples $\theta_1, \ldots, \theta_n \sim (\theta|y_1, ..., y_m)$
2. convert $\theta_1, \ldots, \theta_n$ to $\gamma_1, \ldots, \gamma_n$
3. compute properties of $p(\gamma|y_1, ..., y_m)$ using $\gamma_1, \ldots, \gamma_n$

Example. Consider making comparisons between two properties of a distribution. For example a simple contrast, $\gamma = \theta_1 - \theta_2$, or a more complicated function of θ_1, and θ_2 such as $\gamma = \log(\frac{\theta_1}{\theta_2})$. In a classical setting computing distributions, and associated asymptotic properties can be challenging and require large sample approximations. However, this is simple in a Bayesian framework using the same prescription as above.

Posterior Predictive Distribution

Recall the posterior predictive distribution $p(y^*|y_1, ..., y_n)$ is the predictive distribution for an upcoming data point given the observed data.

How does θ factor into this equation?

$$p(y*|y_1, ..., y_n) = \int p(y*|\theta)p(\theta|y_1, ..., y_n)d\theta \tag{13}$$

Often the predictive distribution is hard to sample from, so a two-step procedure is completed instead.

1. sample θ_i from $p(\theta|y_1, ..., y_n)$
2. sample $y*_i$ from $p(y*_i|\theta_i)$

Similar ideas extend to posterior predictive model checking, which we will return to after studying Bayesian regression.

The Normal Model

A random variable Y is said to be normally distributed with mean θ and variance σ^2 if the density of Y is:

$$p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right]$$

Key points about the normal distribution:

- The distribution is symmetric about θ, and the mode, median and mean are all equal to θ
- about 95% of the mass resides within two standard deviations of the mean
- if $X \sim N(\mu,\tau^2)$ and $Y \sim N(\theta,\sigma^2)$ and X and Y are independent, then $aX + bY \sim N(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)$
- the `rnorm, dnorm, pnorm` and `qnorm` commands in `R` are very useful, but they take σ as their argument not σ^2, so be careful.

Inference for θ, conditional on σ^2

When sigma is known, we seek the posterior distribution of $p(\theta|y_1,....,y_n,\sigma^2)$. A conjugate prior, $p(\theta|sigma^2)$ is of the form:

$$\begin{aligned} p(\theta|y_1,...,y_n,\sigma^2) &\propto p(\theta|\sigma^2) \times \exp\left[-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\right] \\ &\propto \exp\left[c_1(\theta-c_2)^2\right] \end{aligned}$$

thus a conjugate prior for $p(\theta|y_1,...,y_n,\sigma^2)$ is from the normal family of distributions. **Note: verifying this will most likely be a homework problem**

Now consider a prior distribution $p(\theta|\sigma^2) \sim N(\mu_0,\tau_0^2)$ and compute the posterior distribution.

$$\begin{aligned} p(\theta|y_1,\ldots,y_n,\sigma^2) &\propto p(y_1,\ldots,y_n|\theta,\sigma^2)p(\theta|\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\} \times \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\right\} \\ &\quad \text{now combine terms with the powers of } \theta \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu_0}{\tau_0^2}+\frac{\sum y_i}{\sigma^2}\right) + c(y_1,\ldots,y_n,\mu,\sigma^2,\tau_0^2)\right]\right\} \\ &\quad \text{Note from here we could complete the square} \end{aligned}$$

However, there is a shortcut here that you will probably do *approximately* 50 times over the course of this class. Note if $\theta \sim N(E, V)$ then

$$p(\theta) \propto \exp\left[-\frac{1}{2V}(\theta - E)^2\right] \tag{14}$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{\theta^2}{V} - \frac{2\theta E}{V} + c(E, V)\right)\right] \tag{15}$$

Hence from above, the variance of the distribution is the reciprocal of the term with θ^2. That is:

$$V[\theta] = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

Similarly the term associated with 2θ is E/V, so the expectation is this term times the variance. So the expectation is calculated as:

$$E[\theta] = \left(\frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}\right)\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

Notes about the posterior and predictive distributions

- It is common to reparameterize the variance using the inverse, which is known as the precision. Then:
  - $\tilde{\sigma}^2 = 1/\sigma^2$ = sampling precision
  - $\tilde{\tau}_0^2 = 1/\tau_0^2$ = prior precision
  - $\tilde{\tau}_n^2 = 1/\tau_n^2$ = posterior precision, where τ_n^2 is the posterior variance

  Now the posterior precision (i.e. how close the data are to θ) is a function of the prior precision and information from the data: $\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$

- The posterior mean can be reparameterized as a weighted average of the prior mean and the sample mean.

$$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y},$$

  where μ_n is the posterior mean and $\bar{y}$ is the sample mean.

- The predictive distribution of $p(y * |\sigma^2, y_1, \ldots, y_n) \sim N(\mu_n, \tau_n^2 + \sigma)$. This will be a homework problem.

Week 5: Sept 26 - Sept 30

Joint inference for mean and variance in normal model

Thus far we have focused on Bayesian inference for settings with one parameter. Dealing with multiple parameters is not fundamentally different as we use a joint prior $p(\theta_1, \theta_2)$ and use the same mechanics with Bayes rule.

In the normal case we seek the posterior:

$$p(\theta, \sigma^2|y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\theta, \sigma^2)p(\theta, \sigma^2)$$

Recall that $p(\theta, \sigma^2)$ can be expressed as $p(\theta|\sigma^2)p(\sigma^2)$. For now, let the prior on θ the mean term be:

$$p(\theta|\sigma^2) \sim N(\mu_0, \sigma^2/\kappa_0).$$

Then μ_0 can be interpreted as the mean and κ_0 corresponds to the 'hypothetical' number of prior observations. A prior on σ^2 is still needed, a required property for this prior is the the support of $\sigma^2 = (0, \infty)$. A popular distribution with this property is the Gamma distribution. Unfortunately this is not conjugate (or semi-conjugate) for the variance. It turns out that the gamma distribution is conjugate for the precision term $\phi = 1/\sigma^2$, which many Bayesians will use. This implies that the inverse gamma distribution can be used as a prior for σ^2.

For now, set the prior on the precision term $(1/\sigma^2)$ to a gamma distribution. For interpretability this is parameterized as:

$$1/\sigma^2 \sim gamma(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$$

Using this parameterization:

- $E[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$
- mode$[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1}$,
- $Var[\sigma^2]$ is decreasing in ν_0

The nice thing about this parameterization is that σ_0^2 can be interpreted as the sample variance from ν_0 prior samples.

Implementation

Use the following prior distributions for θ and σ^2:

$$\begin{aligned} 1/\sigma^2 &\sim gamma(\nu_0/2, \nu_0\sigma_)^2/2) \\ \theta|\sigma^2 &\sim N(\mu_0, \sigma^2/\kappa_0) \end{aligned}$$

and the sampling model for Y

$$Y_1, \ldots, Y_n|\theta, \sigma^2 \sim i.i.d.\ normal(\theta, \sigma^2).$$

Now the posterior distribution can also be decomposed in a similar fashion to the prior such that:

$$p(\theta, \sigma^2|y_1, \ldots, y_n) = p(\theta|\sigma^2, y_1, \ldots, y_n)p(\sigma^2|y_1, \ldots, y_n).$$

Using the results from the case where σ^2 was known, we get that:

$$\theta|y_1, \ldots, y_n, \sigma^2 \sim normal(\mu_n, \sigma^2\kappa_n),$$

where $\kappa_n = \kappa_0 + n$ and $\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$. Note that this distribution still depends on σ^2 which we do not know.

The **marginal posterior** distribution of σ^2 integrates out θ

$$\begin{aligned} p(\sigma^2|y_1, \ldots, y_n) &\propto p(\sigma^2)p(y_1, \ldots, y_n|\sigma^2) \\ &= p(\sigma^2) \int p(y_1, \ldots, y_n|\theta, \sigma^2)p(\theta|\sigma^2)d\theta \end{aligned}$$

It turns out (**HW?**) that:

$$1/\sigma^2|y_1, \ldots, y_n \sim gamma(\nu_n/2, \nu_n\sigma_n^2/2),$$

where $\nu_n = \nu_0 + n$, $\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right\}$, and $s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$. Again the interpretation is that ν_0 is the prior sample size for σ_0^2.

Posterior Sampling

Now we seek to create draws from the **joint posterior distribution** $p(\theta, \sigma^2|y_1, \ldots, y_n)$ and the **marginal posterior distributions** $p(\theta|y_1, \ldots, y_n)$ and $p(\sigma^2|y_1, \ldots, y_n)$. Note the marginal posterior distributions would be used to calculate quantities such as $Pr[\theta > 0|y_1, \ldots, y_n]$.

Using a Monte Carlo procedure, we can simulate samples from the joint posterior using the following algorithm.

1. Simulate $\sigma_i^2 \sim inverse - gamma(\nu_n/2, \sigma_n^2\nu_n/2)$
2. Simulate $\theta_i \sim normal(\mu_n, \sigma_i^2/\kappa_n)$
3. Repeat m times.

Note that each pair $\{\sigma_i^2, \theta_i\}$ is a sample from the joint posterior distibution and that $\{\sigma_1^2, \ldots, \sigma_m^2\}$ and $\{\theta_1, \ldots, \theta_m\}$ are samples from the respective marginal posterior distributions.
The R code for this follows as:

```
 #### Posterior Sampling with Normal Model
set.seed(09222016)
# true parameters from normal distribution
sigma.sq.true <- 1
theta.true <- 0

# generate data
num.obs <- 100
y <- rnorm(num.obs,mean = theta.true, sd = sqrt(sigma.sq.true))

# specify terms for priors
nu.0 <- 1
sigma.sq.0 <- 1
mu.0 <- 0
kappa.0 <- 1

# compute terms in posterior
kappa.n <- kappa.0 + num.obs
nu.n <- nu.0 + num.obs
s.sq <- var(y) #sum((y - mean(y))^2) / (num.obs - 1)
sigma.sq.n <- (1 / nu.n) * (nu.0 * sigma.sq.0 + (num.obs - 1) *
     s.sq + (kappa.0*num.obs)/kappa.n * (mean(y) - mu.0)^2)
mu.n <- (kappa.0 * mu.0 + num.obs * mean(y)) / kappa.n

# simulate from posterior
#install.packages("LearnBayes")
library(LearnBayes) # for rigamma
num.sims <- 10000
```

```
sigma.sq.sims <- theta.sims <- rep(0,num.sims)
for (i in 1:num.sims){
  sigma.sq.sims[i] <- rigamma(1,nu.n/2,sigma.sq.n*nu.n/2)
  theta.sims[i] <- rnorm(1, mu.n, sqrt(sigma.sq.sims[i]/kappa.n))
}

library(grDevices) # for rgb
plot(sigma.sq.sims,theta.sims,pch=16,col=rgb(.1,.1,.8,.05),
     ylab=expression(theta), xlab=expression(sigma[2]),main='Joint Posterior')
points(1,0,pch=14,col='black')
hist(sigma.sq.sims,prob=T,main=expression('Marginal Posterior of' ~ sigma[2]),
     xlab=expression(sigma[2]))
abline(v=1,col='red',lwd=2)
hist(theta.sims,prob=T,main=expression('Marginal Posterior of' ~ theta),
     xlab=expression(theta))
abline(v=0,col='red',lwd=2)
```

It is important to note that the prior structure is very specific in this case, where $p(\theta|\sigma^2)$ is a function of σ^2. In most prior structures this type of conditional sampling scheme is not as easy as this case and we need to use Markov Chain Monte Carlo methods.

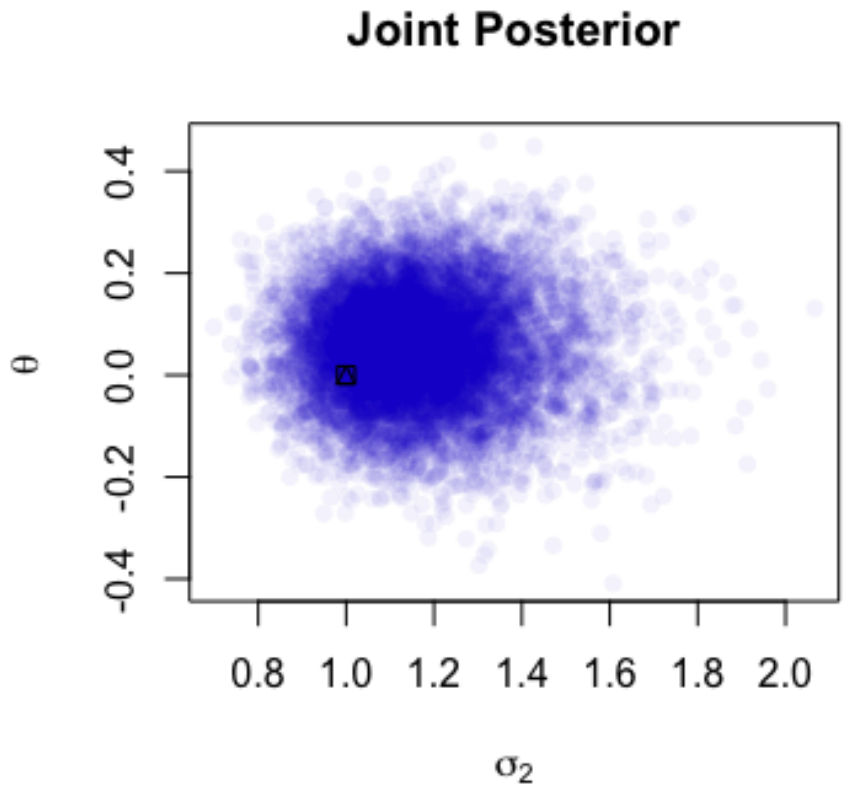
Posterior Sampling with the Gibbs Sampler

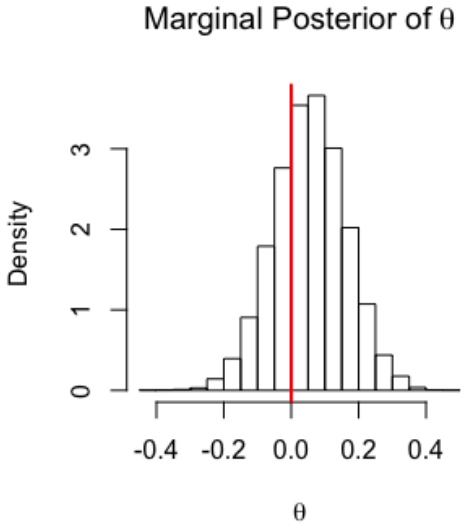
In the previous section we modeled the uncertainty in θ as a function of σ^2, where $p(\theta|\sigma^2) = N(\mu_0, \sigma^2/\kappa_0)$. In some situations this makes sense, but in others the uncertainty in θ may be specified independently from σ^2. Mathematically, this translates to $p(\sigma^2, \theta) = p(\theta) \times p(\sigma^2)$. A common *semiconjugate* set of prior distributions is:

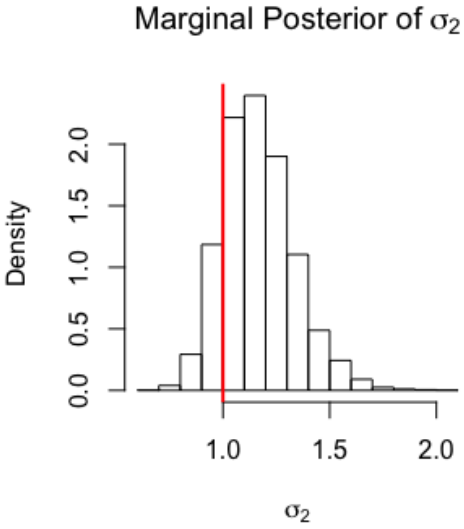
$$\begin{aligned} \theta &\sim normal(\mu_0, \tau_0^2) \\ 1/\sigma^2 &\sim gamma(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

Note this prior on $1/\sigma^2$ is equivalent to saying $p(\sigma^2) \sim InvGamma(\nu_0/2, \nu_0\sigma_0^2/2)$.
Now when $\{Y_1, \ldots, Y_n|\theta, \sigma^2\} \sim normal(\theta, \sigma^2)$ then $\theta|\sigma^2, y_1, \ldots, y_n \sim Normal(\mu_n, \tau_n^2)$.

$$\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \qquad \text{and} \qquad \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$


Joint Posterior
θ
-0.4
-0.2
0.0
0.2
0.4
0.8
1.0
1.2
1.4
1.6
1.8
2.0
σ2


Marginal Posterior of θ
Density
0
1
2
3
-0.4
-0.2
0.0
0.2
0.4
θ


Marginal Posterior of σ2
Density
0.0
0.5
1.0
1.5
2.0
1.0
1.5
2.0
σ2

In the conjugate case where τ_0^2 was proportional to σ^2, samples from the joint posterior can be taken using the Monte Carlo procedure demonstrated before. However, when τ_0^2 is not proportional to σ^2 the marginal density of $1/\sigma^2$ is not a gamma distribution or another named distribution that permits easy sampling.

Suppose that you know the value of θ. Then the conditional distribution of $\tilde{\sigma}^2 = (1/\sigma^2)$ is:

$$\begin{aligned} p(\tilde{\sigma}^2|\theta, y_1, \ldots y_n) &\propto p(y_1, \ldots, y_n|\theta, \tilde{\sigma}^2)p(\tilde{\sigma}^2) \\ &\propto \left((\tilde{\sigma}^2)^{n/2} \exp\left\{ -\tilde{\sigma}^2 \sum_{i=1}^{n} (y_i - \theta)^2/2 \right\} \right) \times \left((\tilde{\sigma}^2)^{\nu_0/2-1} \exp\left\{ -\tilde{\sigma}^2 \nu_0 \sigma_0^2/2 \right\} \right) \\ &\propto (\tilde{\sigma}^2)^{(\nu_0+n)/2-1} \exp\left\{ -\tilde{\sigma}^2 \times \left\langle \nu_0\sigma_0^2 + \sum (y_i - \theta)^2 \right\rangle /2 \right\} \end{aligned}$$

which is the kernel of a gamma distribution. So $\sigma^2|\theta, y_1, \ldots, y_n \sim InvGamma(\nu_n/2, \nu_n\sigma_n^2(\theta)/2)$, where $\nu_n = \nu_0 + n$, $\sigma_n^2(\theta) = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + ns_n^2(\theta)]$ and $s_n^2(\theta) = \sum(y_i - \theta)^2/n$ the unbiased estimate of σ^2 if θ were known.

Now can we use the full conditional distributions to draw samples from the joint posterior?

Suppose we had $\sigma^{2(1)}$, a single sample from the marginal posterior distribution $p(\sigma^2|y_1, \ldots, y_n)$. Then we could sample:

$$\theta^{(1)} \sim p(\theta|y_1, \ldots, y_n, \sigma^{2(1)})$$

and $\{\theta(1), \sigma^{2(1)}\}$ would be a sample from the joint posterior distribution $p(\theta, \sigma^2|y_1, \ldots, y_n)$. Now using $\theta(1)$ we can generate another sample of σ^2 from

$$\sigma^{2(2)} \sim p(\sigma^2|y_1, \ldots, y_n, \theta^{(1)}).$$

This sample $\{\theta^{(1)}, \sigma^{2(2)}\}$ would also be a sample from the joint posterior distribution. This process follows iteratively. However, we don't actually have $\sigma^{2(1)}$.

Gibbs Sampler

The distributions $p(\theta|y_1, \ldots, y_n, \sigma^2)$ and $p(\sigma^2|y_1, \ldots, y_n, \theta)$ are known as the full conditional distributions, that is they condition on all other values and parameters. The Gibbs sampler uses these full conditional distributions and the procedure follows as:

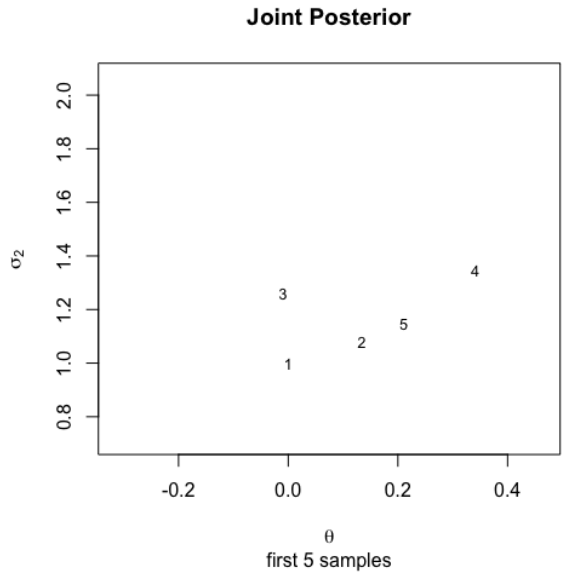
1. sample $\theta(j+1) \sim p(\theta|\tilde{\sigma}^{2(j)}, y_1, \ldots, y_n)$;
2. sample $\tilde{\sigma}^{2(j+1)} \sim p(\tilde{\sigma}^2|\theta^{(j+1)}, y_1, \ldots, y_n)$;
3. let $\phi^{(s+1)} = \{\theta^{(s+1)}, \tilde{\sigma}^{2(s+1)}\}$.

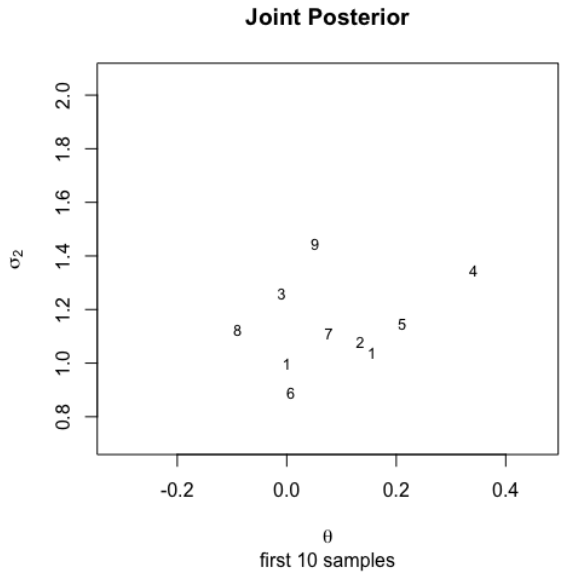
The code and R output for this follows.

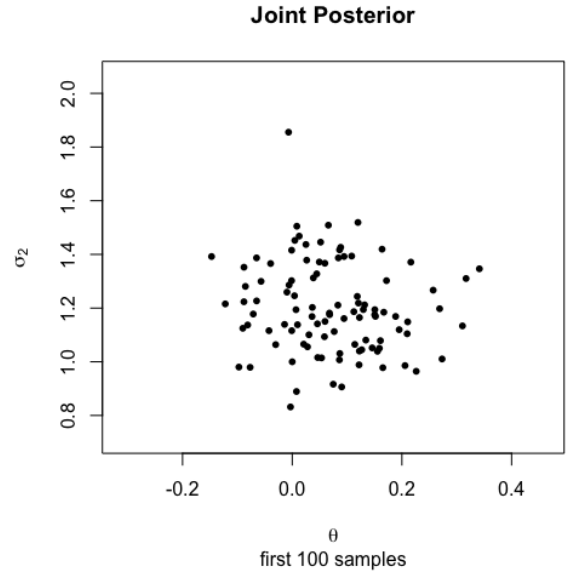
```
######### First Gibbs Sampler
set.seed(09222016)
### simulate data
num.obs <- 100
mu.true <- 0
sigmasq.true <- 1
y <- rnorm(num.obs,mu.true,sigmasq.true)
mean.y <- mean(y)
var.y <- var(y)
library(LearnBayes) # for rigamma
### initialize vectors and set starting values and priors
num.sims <- 10000
Phi <- matrix(0,nrow=num.sims,ncol=2)
Phi[1,1] <- 0 # initialize theta
Phi[1,2] <- 1 # initialize (sigmasq)
mu.0 <- 0
tausq.0 <- 1
nu.0 <- 1
sigmasq.0 <- 1

for (i in 2:num.sims){
  # sample theta from full conditional
  mu.n <- (mu.0 / tausq.0 + num.obs * mean.y / Phi[(i-1),2]) / (1 / tausq.0 + num.obs / Phi[(i-1),2] )
  tausq.n <- 1 / (1/tausq.0 + num.obs / Phi[(i-1),2])
  Phi[i,1] <- rnorm(1,mu.n,sqrt(tausq.n))

  # sample (1/sigma.sq) from full conditional
  nu.n <- nu.0 + num.obs
  sigmasq.n.theta <- 1/nu.n*(nu.0*sigmasq.0 + sum((y - Phi[i,1])^2))
  Phi[i,2] <- rigamma(1,nu.n/2,nu.n*sigmasq.n.theta/2)
}

# plot joint posterior
plot(Phi[1:5,1],1/Phi[1:5,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=c('1','2','3','4','5'),cex=.8,
     ylab=expression(sigma[2]), xlab = expression(theta), main='Joint Posterior',sub='first 5 samples')

plot(Phi[1:10,1],1/Phi[1:10,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=as.character(1:15),cex=.8,
     ylab=expression(sigma[2]), xlab = expression(theta), main='Joint Posterior',sub='first 10 samples')

plot(Phi[1:100,1],1/Phi[1:100,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=16,col=rgb(0,0,0,1),cex=.8,
     ylab=expression(sigma[2]), xlab = expression(theta), main='Joint Posterior',sub='first 100 samples')

plot(Phi[,1],1/Phi[,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=16,col=rgb(0,0,0,.25),cex=.8,
     ylab=expression(sigma[2]), xlab = expression(theta), main='Joint Posterior',sub='all samples')
points(0,1,pch='X',col='red',cex=2)

# plot marginal posterior of theta
hist(Phi[,1],xlab=expression(theta),main=expression('Marginal Posterior of ' ~ theta),probability=T)
abline(v=mu.true,col='red',lwd=2)
# plot marginal posterior of sigmasq
hist(Phi[,2],xlab=expression(sigma[2]),main=expression('Marginal Posterior of ' ~ sigma[2]),probability=T)
abline(v=sigmasq.true^2,col='red',lwd=2)
```

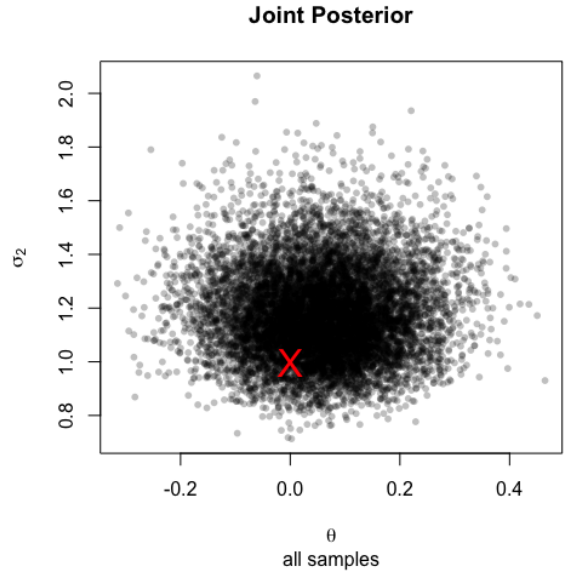
```
# plot trace plots
plot(Phi[,1],type='l',ylab=expression(theta), main=expression('Trace plot for ' ~ theta))
abline(h=mu.true,lwd=2,col='red')
plot(Phi[,2],type='l',ylab=expression(sigma[2]), main=expression('Trace plot for ' ~ sigma[2]))
abline(h=sigmasq.true^2,lwd=2,col='red')

# compute posterior mean and quantiles
colMeans(Phi)
apply(Phi,2,quantile,probs=c(.025,.975))
```

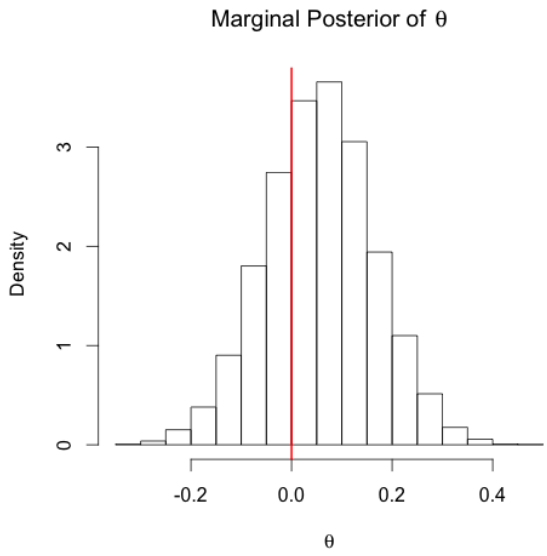
So what do we do about the starting point? We will see that given a reasonable starting point the algorithm will converge to the true posterior distribution. Hence the first (few) iterations are regarded as the burn-in period and are discarded (as they have not yet reached the true posterior).

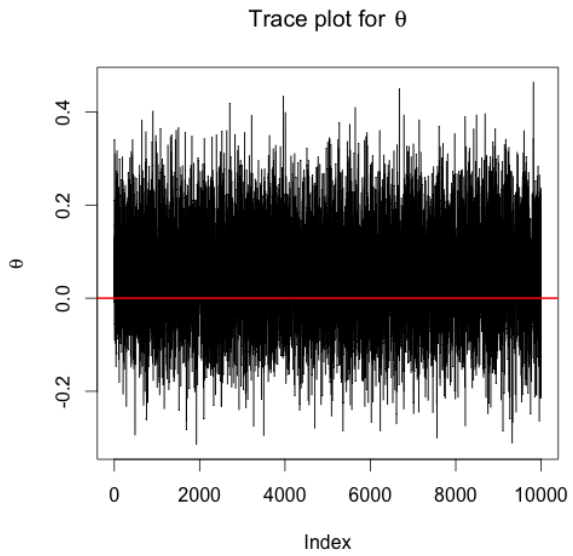

Joint Posterior
σ2
θ
first 5 samples

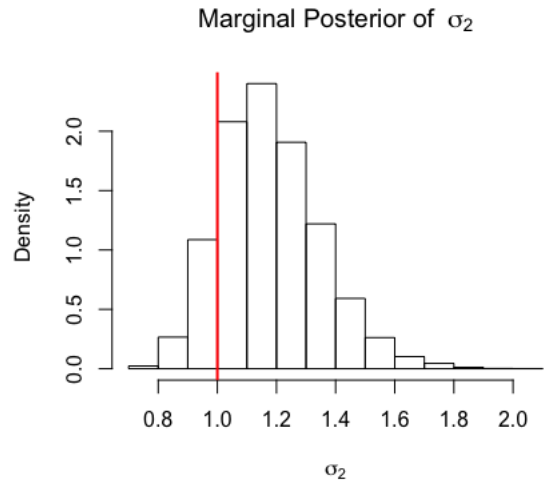

Joint Posterior
σ2
θ
first 10 samples

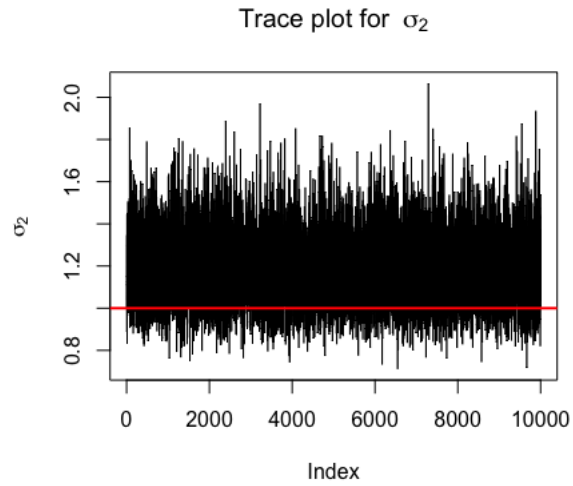

Joint Posterior
σ2
θ
first 100 samples


Joint Posterior
σ2
θ
all samples


Marginal Posterior of θ
Density
θ
0
1
2
3
-0.2
0.0
0.2
0.4


Trace plot for θ
θ
Index
-0.2
0.0
0.2
0.4
0
2000
4000
6000
8000
10000


Marginal Posterior of σ2
Density
σ2
0.0
0.5
1.0
1.5
2.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0


Trace plot for σ2
σ2
Index
0.8
1.2
1.6
2.0
0
2000
4000
6000
8000
10000

Week 6: Oct 3 - Oct 7

More on the Gibbs Sampler

The algorithm previously detailed is called the *Gibbs Sampler* and generates a dependent sequence of parameters $\{\phi_1, \phi_2, \ldots, \phi_n\}$. This is in contrast to the Monte Carlo procedure we previously detailed, including the situation where $p(\theta|\sigma^2) \sim N(\mu_0, \sigma^2/\kappa_0)$.

The Gibbs Sampler is a basic Markov Chain Monte Carlo (MCMC) algorithm. A Markov chain is a stochastic process where the current state only depends on the previous state. Formally

$$Pr(X_n = x_n|X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0) = Pr(X_n = x_n|X_{n-1} = x_{n-1})$$

Depending on the class interests, we may return to talk more about the theory of MCMC later in the course, but the basic ideas are:

$$Pr(\theta^{(j)} \in A) \rightarrow \int_A p(\phi)d\phi \quad \text{as } j \rightarrow \infty.$$

That is the sampling distribution of the draws from the MCMC algorithm approach the desired target distribution (generally a posterior in Bayesian statistics) as the number of samples j goes to infinity. The is not dependent on the starting values of $\phi^{(0)}$, but poor starting values will take longer for convergence. Note this will be more problematic when we consider another MCMC algorithm, the Metropolis-Hastings sampler. Given the equation above, for most functions $g(.)$:

$$\frac{1}{J}\sum_{j=1}^{J} g(\phi^{(j)}) \rightarrow E[g(\phi)] = \int g(\phi)p(\phi)d\phi \quad \text{as } J \rightarrow \infty.$$

Thus we can approximate expectations of functions of ϕ using the sample average from the MCMC draws, similar to our Monte Carlo procedures presented earlier.

Estimation vs. Approximation

There are two elements of a Bayesian data analysis:

1. *Model Specification:* a sampling model $p(y_1, \ldots, y_n|\phi)$ is specified. Where ϕ could be a high dimensional parameter set.

2. *Prior Specification:* a probability distribution $p(\phi)$ is specified.
   Once these are specified and the data have been gathered, the posterior distribution $p(\phi|y_1, \ldots, y_n)$ is fully determined. It is exactly:

$$p(\phi|y_1, \ldots, y_n) = \frac{p(\phi)p(y_1, \ldots, y_n|\phi)}{p(y_1, \ldots, y_n)}$$

   Excluding posterior model checking, all that remains is to summarize the posterior.

3. *Posterior Summary:* A description of the posterior distribution $p(\phi|y_1, \ldots, y_n)$, typically using intervals, posterior means, and predictive probabilities.

For most models we have discussed thus far, $p(\phi|y_1, \ldots, y_n)$ is known in closed form or easy to sample from using Monte Carlo procedures. However, in more sophisticated settings, $p(\phi|y_1, \ldots, y_n)$ is complicated, and hard to write down or sample from. In these cases, we study $p(\phi|y_1, \ldots, y_n)$ by looking at MCMC samples. Thus, Monte Carlo and MCMC sampling algorithms:

- are not models,
- they do not generate "more information" than is in $y_1, \ldots, y_n$ and $p(\phi)$
- they are simply 'ways of looking at' $p(\phi|y_1, \ldots, y_n)$

For example, if we have Monte Carlo samples $\phi^{(1)}, \ldots, \phi^{(J)}$ that are approximate draws from $p(\theta|y_1, dots, y_n)$, then these sample help describe $p(\phi|y_1, \ldots, y_n)$, for example:

- $\frac{1}{J}\sum \phi^{(j)} \approx \int \phi p(\phi|y_1, \ldots, y_n)d\phi$
- $\frac{1}{J}\sum \delta(\phi^{(j)} \leq c) \approx Pr(\phi \leq c|y_1, \ldots, y_n) = \int_{-\infty}^{c} p(\phi|y_1, \ldots, y_n)d\phi$

To keep the distinction between *estimation* and *approximation* clear, commonly *estimation* is used do describe how we use $p(\phi|y_1, \ldots, y_n)$ to make inferences about ϕ and use *approximation* to describe the use of Monte Carlo (including MCMC) procedures to approximate integrals.

MCMC Exercise

The purpose of Monte Carlo or MCMC algorithms is to generate a sequence of sample draws $\{\phi^{(1)}, \ldots, \phi^{(J)}\}$ such that we can approximate

$$\frac{1}{J}\sum_{j=1}^{J} g(\phi^{(j)}) \approx \int g(\phi)p(\phi)d\phi,$$

where $g(.)$ could be an expectation, quantile, or other property of the posterior distribution. In order for this approximation to be 'good' the empirical distribution of the simulated sequence $\{\phi^{(1)}, \ldots, \phi^{(J)}\}$ needs

to lok like the target distribution $p(\phi)$. Note the target distribution is typically a posterior, but this notation allows $p(\phi)$ to be a generic distribution function.

Monte Carlo procedures allow us to generate independent sample from the target distribution; hence, these samples are representative of the target distribution. The only question is how many sample to we need to attain a specified level of precision (usually CLT theory for approximating expectations). However, with MCMC samples we generate dependent samples and all we are guaranteed theoretically is:

$$\lim_{J\to\infty} Pr(\phi^{(j)} \in A) = \int_A p(\phi)d\phi$$

In other words, our posterior samples from the MCMC algorithm will *eventually* converge. However, there is no guarantee that the convergence can even be acheived in a practical amount of computing time.

Example. An example that is notoriously difficult to acheive convergence is high-dimensional multimodal settings. Similar to optimization algorithms, MCMC samplers can 'get stuck' in a single mode.

Exercise. Will be given as an in class lab/quiz Consider the mixture distribution described on p. 99 (Hoff). This distribution is a joint probability distribution of a discrete variable $\delta = \{1, 2, 3\}$, denoting which mixture component the mass comes from and a continuous variable θ. The target density is $\{Pr(\delta = 1), Pr(\delta = 2), Pr(\delta = 3)\} = (.45, .10, .45)$ and $p(\theta|\delta = i) \sim N(\theta; \mu_i, \sigma_i^2)$ where $\{\mu_1, \mu_2, \mu_3\} = (-3, 0, 3)$ and $\sigma_i^2 = 1/3$ for $i \in \{1, 2, 3\}$.

1. Generate 1000 samples of θ from this distribution using a Monte Carlo procedure. Hint: first generate $\delta^{(i)}$ from the marginal distribution $p(\delta)$ and then generate $\theta^{(i)}$ from $p(\theta|\delta)$ Plot your samples in a histogram form and superimpose a curve of the density function. Comment on your samples, do they closely match the true distribution?

2. Next, generate samples from a Gibbs sampler using the full conditional distributions of θ and δ. You already know the form of the full conditional for θ from above. The full conditional distribution for δ is given below:

$$Pr(\delta = d|\theta) = \frac{Pr(\delta = d) \times p(\theta|\delta = d)}{\sum_{d=1}^{3} Pr(\delta = d) \times p(\theta|\delta = d)}$$

   Hint: for $p(\theta|\delta = d)$ evaluate θ from a normal distribution with parameters $\{\mu_d, \sigma_d^2\}$. Intialize θ at 0.

   (a) Generate 100 samples using this procedure. Plot your samples as a histogram with the true density superimposed on the plot. Also include a plot of your θ value on the y-axis and the iteration number on the x-axis. This is called a trace plot, and allows your to visualize the

movement of your MCMC *particle*. Comment on how close your samples match the true density. What does the trace plot reveal about the position of θ over time (the iterations)? Does the proportion of the time the sample spends in each state (δ) match the true probabilities?

(b) Repeat for 1000 samples.

(c) Repeat for 10000 samples.

3. Now repeat part 2, but instead initialize θ at 100. How does this change the results from part 2?

When complete, turn your code in as a R markdown document for quiz 3.

MCMC Diagnostics

A useful way to think about an MCMC sampler is that there is a *particle* moving through and exploring the parameter space. For each region, or set, A the particle needs to spend time proportional to the target probability, $\int_A p(\phi)d\phi$ Consider the three modes from the exercise conducted in class and denote these three modes as A_1, A_2, A_3. Given the weights on the mixture components the particle should spend more time in A_1 and A_3 than A_2. However, if the particle was initialized in A_2 we'd hope that the number of iterations are large enough that:

1. The particle moves out of A_2 and into higher probability regions and
2. the particle moves between A_1, A_2, and A_3.

The technical term associated with item 1 is *stationarity*, which means the chain has conveged to the target distribution. For the models we have seen thus far, convergence happens quite rapidly, but we will look at this in more depth later on. The second item is focused on the speed the particle moves through the target distribution, this is referred to as *mixing*. An independent sampler like the Monte Carlo procedures we have seen have perfect mixing as each sample is independently drawn from the target distribution. The MCMC samples can be highly correlated and tend to get stuck in certain regions of the space.

People often quantify mixing properties of MCMC samples using the idea of effective sample size. To understand this, first consider the variance of independent Monte Carlo samples:

$$Var_{MC}[\bar{\phi}] = \frac{Var[\phi]}{J},$$

where $\bar{\phi} = \sum_{j=1}^{J} \phi^{(j)}/J$. The Monte Carlo variance is controlled by the number of samples obtained from the algorithm. In a MCMC setting, consecutive samples $\phi^{(j)}$ and $\phi^{(j+1)}$ are not independent, rather

they are usually positively correlated. Once stationarity has been acheived, the variance of the MCMC algorithm can be expressed as:

$$Var_{MCMC}[\phi] = \cdots = Var_{MC}[\bar{\phi}] + \frac{1}{S^s}\sum_{j \neq k} E[(\phi^{(j)} - \phi_0)(\phi^{(k)} - \phi_0)],$$

where ϕ_0 is the true value of the integral, typically $E[\phi]$. Now if two consecutive samples are highly correlated the variance of the estimator will be much larger than that of an Monte Carlo procedure with the same number of iterations. This is captured in the idea of the **effective sample**. The effective sample size is computed such that:

$$Var_{MCMC}[\bar{\phi}] = \frac{Var[\phi]}{S_{eff}},$$

where S_{eff} can be interpreted as the number of independent Monte Carlo samples necessary to give the same precision as the MCMC samples. Note that the R function `effectiveSize` in the "coda" package will calculate the effective sample size of MCMC output.

We will talk more about MCMC diagnostics after introducing the Metropolis-Hastings algorithm later in class, but the general procedure is:

1. Run multiple chains from different starting points
2. Assess the similarity of different chains, ideally visually and/or with a test statistic

An easy solution, especially in the context of Gibbs Sampling is to look at trace plots and histograms of marginal posterior distributions. In conjunction with ESS (Effective Sample Size) calculations this usually gives a good sense of convergence. In other situations, combining visuals displays (trace plots) with other statisics Gelman's R statistic or QDE is a good strategy.

The big picture idea with MCMC, is that we want to guarantee that our algorithm has:

1. Reached stationarity, that is converged to the true target distribution and
2. is efficiently mixing, that is the particle can effectively sweep through the distribution without getting stuck in regions.

Posterior Model Checks on Normal Model

Exercise. Consider a similar scenario to the code for the first Gibbs sampler. Again 100 data points have been generated.

1. A histogram of the data is shown later as figure (a). What are your thoughts about this data?

2. Now assume you used your MCMC code and came up with figures (b) - (e). Comment on the convergence and the marginal posterior distributions.

3. As a final check you decide to use your MCMC samples to compute the posterior predictive distribution, $p(y^*|y_1, \ldots, y_n)$. Computationally this can be acheived by using each pair $\{\theta^{(i)}, \sigma^{2(i)}\}$ and then simulating from $N(y^*; \theta^{(i)}, \sigma^{2(i)})$. In R this can be done with one line of code:

   ```
   post.pred <- rnorm(num.sims,mean=Phi[,1],sd = sqrt(1/Phi[,2]))
   ```

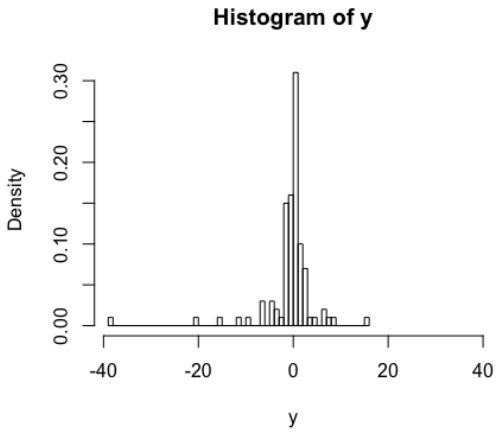
   Now compare your posterior predictive distribution with the observed data. Are you satisfied that the posterior predictive distribution represents the actual data? Why or why not?

We will soon see a trick for handling this case where the data is overdispersed relative to a normal distribution. This is extremely useful in modeling data, as we can use a Gibbs sampler with another parameter to fit the entire class of t-distributions.

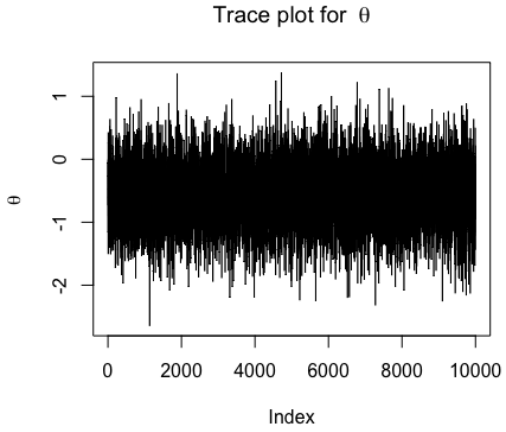
Week 8: Oct 17 - Oct 21

While the normal distribution has two parameters, up to now we have focused on univariate data. Now we will consider multivariate responses from a multivariate normal distribution.

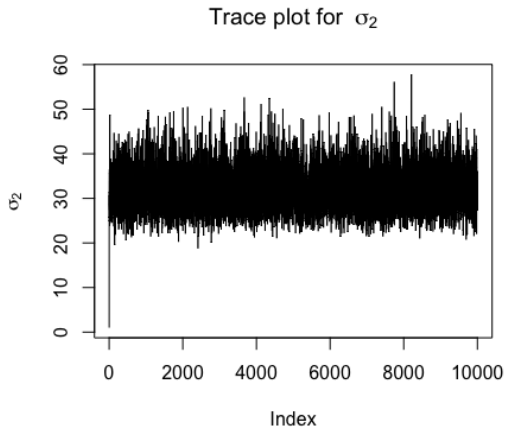
Example. Consider a study addressing colocated environmental variables. For instance, bacteria concentrations and turbidty measurements in a watershed are likely correlated. We are interested in learning not only the mean and variance terms for each variable, but also the correlation structure between the two variables.



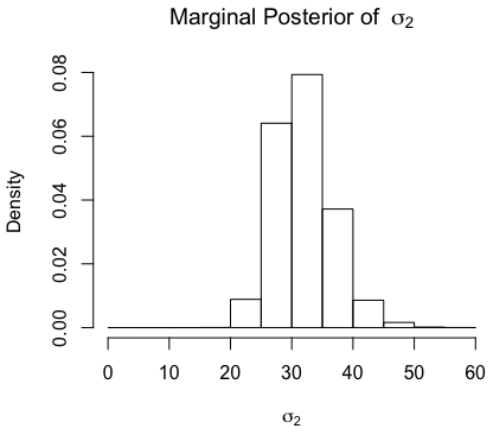
(a) Histogram of the Data



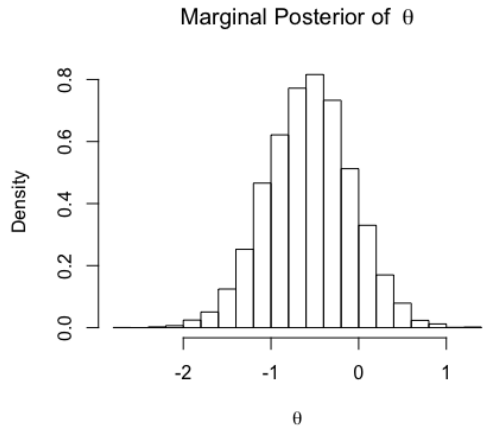
(b) Trace plot for θ

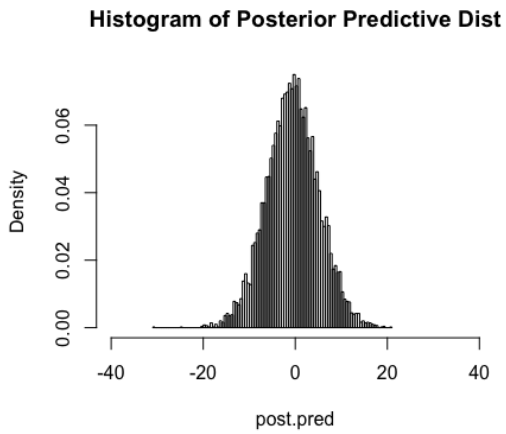


(c) Trace plot for σ^2



(d) Histogram of marginal posterior for σ^2





(e) Histogram of marginal posterior for θ

Multivariate Normal Distribution: The multivariate normal distribution has the following sampling distribution:

$$p(\tilde{y}|\tilde{\theta}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left[-(\tilde{y} - \tilde{\theta})^T \Sigma^{-1} (\tilde{y} - \tilde{\theta}/2\right],$$

where

$$\tilde{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \qquad \tilde{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_p^2 \end{pmatrix}$$

The vector $\tilde{\theta}$ is the mean vector and the matrix Σ is the covariance matrix, where the diagonal elements are the variance terms for observation i and the off diagonal elements are the covariance terms between observation i and j. Marginally, each $y_i \sim N(\theta_i, \sigma_i^2)$.

Linear Algebra Review

- Let A be a matrix, then $|A|$ is the **determinant** of A. For a 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $|A| = ad - bc$. The R command `det(.)` will calculate the determinant of a matrix.
- Let A be a matrix, then A^{-1} is the **inverse** of A such that $AA^{-1} = I_p$ where I_p is the identity matrix of dimension p. The R function `solve(.)` will return the inverse of the matrix. Note that this can be computationally difficult, so whenever possible avoid computing this in every iteration of a sampler.
- Let $\tilde{\theta}$ be a vector of dimension $p \times 1$, $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}$, then the transpose of θ, denoted θ^T is a vector of dimension $1 \times p$. Then $\theta^T = (\theta_1, \theta_2, \cdots, \theta_p)$. The R function `t(.)` will compute the transpose of a vector or matrix.
- Let A be a $n \times p$ matrix and B be a $p \times m$ matrix, then the matrix product, AB will be a $n \times m$ matrix. The element in the i^{th} row and j^{th} column is equal to the vector product of row i of matrix A and column j of matrix B. For vector multiplication in R use the command `%*%`, that is $AB <- A\%*\%B$.
- Let A be a matrix, then the **trace** of A is the sum of the diagonal elements. A usual property of the trace is that $tr(AB) = tr(BA)$. In R the `matrix.trace(.)` function from the matrixcalc

package will return the trace.

- The `mnormt` package in R includes functions `rmnorm, dmnorm` which are multivariate analogs of functions used for a univariate normal distribution.

Priors for multivariate normal distribution

In the univariate normal setting, a normal prior for the mean term was *semiconjugate*. Does the same hold for a multivariate setting? Let $p(\tilde{\theta}) \sim N_p(\tilde{\mu}_0, \Lambda_0)$.

$$\begin{aligned} p(\tilde{\theta}) &= (2\pi)^{-p/2}|\Lambda_0|^{-1/2}\exp\left[-\frac{1}{2}(\tilde{\theta}-\tilde{\mu}_0)^T\Lambda_0^{-1}(\tilde{\theta}-\tilde{\mu_0})\right] \\ &\propto \exp\left[-\frac{1}{2}(\tilde{\theta}-\tilde{\mu}_0)^T\Lambda_0^{-1}(\tilde{\theta}-\tilde{\mu_0})\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\tilde{\theta}^T\Lambda_0^{-1}\tilde{\theta}-\tilde{\theta}^T\Lambda_0^{-1}\tilde{\mu_0}-\tilde{\mu_0}^T\Lambda_0^{-1}\tilde{\theta}\right)\right] \end{aligned}$$

Now combine this with the sampling model, only retaining the elements that contain θ.

$$\begin{aligned} p(\tilde{y}_1,\ldots,\tilde{y}_n|\tilde{\theta},\Sigma) &\propto \prod_{i=1}^{n}\exp\left[-\frac{1}{2}(\tilde{y_i}-\tilde{\theta})^T\Sigma^{-1}(\tilde{y_i}-\tilde{\theta})\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\tilde{y_i}-\tilde{\theta})^T\Sigma^{-1}(\tilde{y_i}-\tilde{\theta})\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\tilde{\theta}^T n\Sigma^{-1}\tilde{\theta}-\tilde{\theta}^T\Sigma^{-1}\sum_{i=1}^{n}\tilde{y}_i-\sum_{i=1}^{n}\tilde{y}_i^T\Sigma^{-1}\tilde{\theta}\right)\right] \end{aligned}$$

Next we find the full conditional distribution for θ, $p(\tilde{\theta}|\Sigma,\tilde{y}_1,\ldots,\tilde{y}_n)$.

$$\begin{aligned} p(\tilde{\theta}|\Sigma,\tilde{y}_1,\ldots,\tilde{y}_n) &\propto \exp\left[-\frac{1}{2}\left(\tilde{\theta}^T n\Sigma^{-1}\tilde{\theta}-\tilde{\theta}^T\Sigma^{-1}\sum_{i=1}^{n}\tilde{y}_i-\sum_{i=1}^{n}\tilde{y}_i^T\Sigma^{-1}\tilde{\theta}+\tilde{\theta}^T\Lambda_0^{-1}\tilde{\theta}-\tilde{\theta}^T\Lambda_0^{-1}\tilde{\mu_0}-\tilde{\mu_0}^T\Lambda_0^{-1}\tilde{\theta}\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\tilde{\theta}^T(n\Sigma^{-1}+\Lambda_0^{-1})\tilde{\theta}-\tilde{\theta}^T(\Sigma^{-1}\sum_{i=1}^{n}\tilde{y}_i+\Lambda_0^{-1}\tilde{\mu}_0)-c\tilde{\theta}\right)\right] \\ &\quad \text{it turns out we can drop the term } c\tilde{\theta} \\ &\propto \exp\left[-\frac{1}{2}\left(\tilde{\theta}^T A\tilde{\theta}-\tilde{\theta}^T B\right)\right] \\ &\quad \text{and we have a similar result to that found earlier for a univariate normal} \end{aligned}$$

The variance (matrix) is A^{-1} and the expectation is $A^{-1}B$. Hence the full conditional follows a multivariate normal distribution with variance $\Lambda_n = \left(n\Sigma^{-1}+\Lambda_0^{-1}\right)^{-1}$ and expectation

= $\tilde{\mu}_n = \left(n\Sigma^{-1} + \Lambda_0^{-1}\right)^{-1}\left(\Sigma^{-1}\sum_{i=1}^n \tilde{y}_i + \Lambda_0^{-1}\tilde{\mu}_0\right)$. Sometimes a uniform prior $p(\tilde{\theta}) \propto \tilde{1}$ is used. In this case the variance and expectation simplify to $V = \Sigma/n$ and $E = \bar{\tilde{y}}$.

Using this semiconjugate prior in a Gibbs sampler we can make draws from the full conditional distribution using `rmnorm(.)` in R. However, we still need to be able to take samples of the covariance matrix Σ to get draws from the joint posterior distribution.

Inverse-Wishart Distribution

A covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_p^2 \end{pmatrix}$ has the variance terms on the diagonal and covariance terms for off diagonal elements. Similar to the requirement that σ^2 be positive, a covariance matrix, Σ must be **positive definite**, such that: $\tilde{x}^T\Sigma\tilde{x} > 0$ for all vectors $\tilde{x}$. With a positive definite matrix, the diagonal elements (which correspond the marginal variances σ_j^2) are greater than zero and it also constrains the correlation terms to be between -1 and 1. A covariance matrix also requires symmetry, so that $Cov(y_i, y_j) = Cov(y_j, y_i)$.

The covariance matrix is closely related to the sum of squares matrix with is given by:

$$\sum_{i=1}^N \tilde{z}_i\tilde{z}_i^T = Z^TZ,$$

where $z_1, ..., z_n$ are $p \times 1$ vectors containing the multivariate response. Thus $\tilde{z}_i\tilde{z}_i^T$results in a $p \times p$ matrix, where

$$\tilde{z}_i\tilde{z}_i^T = \begin{pmatrix} z_{i,1}^2 & z_{i,1}z_{i,2} & \cdots & z_{i,1}z_{i,p} \\ z_{i,2}z_{i,1} & z_{i,2}^2 & \cdots & z_{i,2}z_{i,p} \\ \vdots & & & \vdots \\ z_{i,p}z_{i,1} & z_{i,p}z_{i,2} & \cdots & z_{i,p}^2 \end{pmatrix}$$

Now let the $\tilde{z}_i's$ have zero mean (are centered). Recall that the sample variance is computed as $S^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1) = \sum_{i=1}^n z_i^2/(n-1)$. Similarly the matrix $\tilde{z}_i\tilde{z}_i^T/n$ is the contribution of the i^{th} observation to the sample covariance. In this case:

- $\frac{1}{n}[Z^TZ]_{j,j} = \frac{1}{n}\sum_{i=1}^n z_{i,j}^2 = s_j^2$ That is the diagonal elements of the matrix Z^TZ are an estimate of the marginal sample variances.
- $\frac{1}{n}[Z^TZ]_{j,k} = \frac{1}{n}\sum_{i=1}^n z_{i,j}z_{i,k} = s_{j,k}$ That is the off-diagonal elements of the matrix Z^TZ are estimates of the covariance terms.

If $n > p$ and the $\tilde{z}_i's$ are linearly independent then Z^TZ will be positive definite and symmetric.

Consider the following procedure with a positive integer, ν_0, and a $p \times p$ covariance matrix Φ_0:

1. sample $\tilde{z}_1, \ldots, \tilde{z}_{\nu_0} \sim MVN(\tilde{0}, \Phi_0)$
2. calculate $Z^TZ = \sum_{i=1}^{\nu_0} \tilde{z}_i\tilde{z}_i^T$

then the matrix Z^TZ is a random draw from a **Wishart distribution** with parameters ν_0 and Φ_0. The expectation of Z^TZ is $\nu_0\Phi_0$. The Wishart distribution can be thought of as a multivariate analogue of the gamma distribution. Accordingly, the Wishart distribution is semi-conjugate for the precision matrix (Σ^{-1}); whereas, the Inverse-Wishart distribution is semi-conjugate for the covariance matrix.

The density of the inverse-Wishart distribution with parameters S_0^{-1}, a $p \times p$ matrix and ν_0 ($IW(\nu_0, S_0^{-1})$ is:

$$\begin{aligned} p(\Sigma) &= \left[2^{\nu_0 p/2}\pi^{\binom{p}{n}/2}|S_0|^{-\nu_0/2}\prod_{j=1}^{P}\Gamma([\nu_0+1-j]/2)\right]^{-1} \times \\ & \quad |\Sigma|^{-(\nu_0+p+1)/2} \times \exp[-tr(S_0\Sigma^{-1})/2] \end{aligned}$$

Inverse Wishart Full Conditional Calculations

$$\begin{aligned} p(\Sigma|\tilde{y}_1, \ldots, \tilde{y}_n, \tilde{\theta}) &\propto p(\Sigma) \times p(\tilde{y}_1, \ldots, \tilde{y}_n|\Sigma, \tilde{\theta}) \\ &\propto \left(|\Sigma|^{-(\nu_0+p+1)/2}\exp[-tr(S_0\Sigma^{-1})/2]\right) \times \left(|\Sigma|^{-n/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(\tilde{y}_i-\tilde{\theta})^T\Sigma^{-1}(\tilde{y}_i-\tilde{\theta})]\right) \\ & \quad \text{note that } \sum_{i=1}^{n}(\tilde{y}_i-\tilde{\theta})^T\Sigma^{-1}(\tilde{y}_i-\tilde{\theta}) \text{ is a number so we can apply the trace operator} \\ & \quad \text{using properties of the trace, this equals } tr\left(\sum_{i=1}^{n}(\tilde{y}_i-\tilde{\theta})(\tilde{y}_i-\tilde{\theta})^T\Sigma^{-1}\right) = tr(S_\theta\Sigma^{-1}) \\ \text{so } p(\Sigma|\tilde{y}_1, \ldots, \tilde{y}_n, \tilde{\theta}) &\propto \left(|\Sigma|^{-(\nu_0+p+1)/2}\exp[-tr(S_0\Sigma^{-1})/2]\right) \times \left(|\Sigma|^{-n/2}\exp[-tr(S_\theta\Sigma^{-1}/2)]\right) \\ &\propto \left(|\Sigma|^{-(\nu_0+n+p+1)/2}\exp[-tr([S_0+S_\theta]\Sigma^{-1})/2]\right) \\ \text{thus}\Sigma|\tilde{y}_1, \ldots, \tilde{y}_n, \tilde{\theta} &\sim IW(\nu_0+n, [S_0+S_\theta]^{-1}) \end{aligned}$$

Thinking about the parameters in the prior distribution, ν_0 is the prior sample size and S_0 is the prior residual sum of squares.

Gibbs Sampling for Σ and $\tilde{\theta}$

We now that the full conditional distributions follow as:

$$\begin{aligned}\tilde{\theta}|\Sigma, \tilde{y}_1, \ldots, \tilde{y}_n &\sim MVN(\mu_n, \Lambda_n) \\ \Sigma|\tilde{\theta}, \tilde{y}_1, \ldots, \tilde{y}_n &\sim IW(\nu_n, S_n^{-1}).\end{aligned}$$

Given these full conditional distributions the Gibbs sampler can be implemented as:

1. Sample $\tilde{\theta}^{(j+1)}$ from the full conditional distribution
   (a) compute $\tilde{\mu}_n$ and Λ_n from $\tilde{y}_1, \ldots, \tilde{y}_n$ and $\Sigma^{(j)}$
   (b) sample $\tilde{\theta}^{(j+1)} \sim MVN(\tilde{\mu}_n, \Lambda_n)$. This can be done with `rmnorm(.)` in R.
2. Sample $\Sigma^{(j+1)}$ from its full conditional distribution
   (a) compute S_n from $\tilde{y}_1, \ldots, \tilde{y}_n$ and $\tilde{\theta}^{(j+1)}$
   (b) sample $\Sigma^{(j+1)} \sim IW(\nu_0 + n, S_n^{-1})$

As $\tilde{\mu}_n$ and Λ_n depend on Σ they must be calculated every iteration. Similarly, S_n depends on $\tilde{\theta}$ and needs to be calculated every iteration as well.

Week 9 Oct 24 - Oct 28

Hierarchical Modeling

This chapter focuses on comparison of means across groups and more generally Bayesian hierarchical modeling. Hierarchical modeling is defined by datasets with a multilevel structure, such as:

- patients within hospitals or
- students within school.

The most basic form of this type of data consists of two-levels, groups and individuals within groups.

Recall, observations are exchangeable if $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$. Consider where $Y_1, \ldots, Y_n$ are

test scores from randomly selected students from a given STAT 216 instructor/course. If exchangeability holds for these values, then:

$$\begin{aligned} \phi &\sim p(\phi), \\ Y_1, \ldots, Y_n|\phi &\sim \text{i.i.d. } p(y|\phi). \end{aligned}$$

The exchangeability can be interpreted that the random variables are independent samples from a population with a parameter, ϕ. For instance in a normal model, $\phi = \{\theta, \sigma^2\}$ and the data are conditionally independent from a normal distribution $N(\theta, \sigma^2)$.

In a hierarchical framework this can be extended to include the group number:

$$Y_{1,j}, \ldots, Y_{n_j,j}|\phi_j \sim \text{i.i.d. } p(y|\phi_j).$$

The question now is how to we characterize the information between $\phi_1, \ldots, \phi_m$?

Is it reasonable to assume that the values are independent, that is does the information from ϕ_i tell you anything about ϕ_j?

Now consider the groups as samples from a larger population, then using the idea of exchangeability with group-specific parameters gives:

$$\phi_1, \ldots, \phi_m|\psi \sim \text{i.i.d. } p(\phi|\psi).$$

This is similar to the idea of a random effect model and gives the following hierarchical probability model:

$$\begin{aligned} y_{1,j}, \ldots, y_{n_j,j}|\phi_j &\sim p(y|\phi_j) && \text{(within-group sampling variability)} \\ \phi_1, \ldots, \phi_m|\psi &\sim p(\phi|\psi) && \text{(between-group sampling variability)} \\ \psi &\sim p(\psi) && \text{(prior distribution)} \end{aligned}$$

The distributions $p(y|\phi)$ and $p(\phi|\psi)$ represent sampling variability:

- $p(y|\phi)$ represents variability among measurements within a group and
- $p(\phi|\psi)$ represents sampling variability across groups.

Hierarchical normal model

The hierarchical normal model is often used for modeling differing means across a population.

$$\begin{aligned} \phi_j = \{\theta_j, \sigma^2\}, p(y|\phi_j) &= normal(\theta_j, \sigma^2) \text{ within-group model} \\ \psi = \{\mu, \tau\}, p(\theta_j|\psi) &= normal(\mu, \tau^2) \text{ between-group model} \end{aligned}$$

Note this model specification assumes constant variance for each within-group model, but this assumption can be relaxed.

This model contains three unknown parameters that need priors, we will use the standard semi-conjugate forms:

$$\begin{aligned}\sigma^2 &\sim InvGamma(\nu_0/2, \nu_0\sigma_0^2/2)\\ \tau^2 &\sim InvGamma(\eta_0/2, \eta_0\tau_0^2/2)\\ \mu &\sim normal(\mu_0, \gamma_0^2)\end{aligned}$$

Given these priors, we need to derive the full conditional distributions in order to make draws from the posterior distribution. Note the joint posterior distribution, can be expressed as:

$$\begin{aligned}p(\tilde{\theta}, \mu, \tau^2, \sigma^2|\tilde{y}_1, \ldots, \tilde{y}_n) &\propto p(\mu, \tau^2, \sigma^2) \times p(\tilde{\theta}|\mu, \tau^2, \sigma^2) \times p(\tilde{y}_1, \ldots, \tilde{y}_m|\tilde{\theta}, \mu, \tau^2, \sigma^2)\\ &\propto p(\mu)p(\sigma^2)p(\tau^2) \times \left(\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)\right) \times \left(\prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma^2)\right).\end{aligned}$$

- **Sampling** μ: $p(\mu|-) \propto p(\mu)\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)$. This is a familiar setting with two normal models, hence, the posterior is also a normal distribution.
  - $\mu|- \sim normal\left(\frac{m\bar{\theta}/\tau^2+\mu_0/\gamma_0^2}{m/\tau^2+1/\gamma_0^2}, [m/\tau^2 + 1/\gamma_0^2]^{-1}\right)$
- **Sampling** τ^2: $p(\tau^2|-) \propto p(\tau^2)\prod_{j=1}^{m} p(\theta_j|\mu, \tau^2)$. Again this is similar to what we have seen before.
  - $\tau^2|- \sim InvGamma\left(\frac{\eta_0+m}{2}, \frac{\eta_0\tau_0^2+\sum_j(\theta_j-\mu)^2}{2}\right)$

Now what about $\theta_1, \ldots, \theta_m$?

- **Sampling** $\theta_1, \ldots, \theta_m$. Consider a single θ_j, then $\theta_j|- \propto p(\theta_j|\mu, \tau^2)\prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma^2)$. Again this is the case where we have two normal distributions.
  - $\theta_j|- \sim normal\left(\frac{n_j,\bar{y}_j/\sigma^2+1/\tau^2}{n_j/\sigma^2+1/\tau^2}, [n_j/\sigma^2 + 1/\tau^2]^{-1}\right)$
- **Sampling** σ^2: $p(\sigma^2|-) \propto p(\sigma^2)\prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma)$.
  - $\sigma^2|- \sim InvGamma\left(\frac{1}{2}\left[\nu_0 + \sum_{j=1}^{m} n_j\right], \frac{1}{2}\left[\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2\right]\right).$

Data Example

Consider the dataset outline in Chapter 8, that focuses on math tests scores for students spread across 100 schools. Using the Gibbs sampling procedure described above we can fit this model, code courtesy of textbook.

```
> Y.school.mathscore<-dget("http://www.stat.washington.edu/~hoff/Book/Data/data/Y.school.mathscore")
>
> Y <- Y.school.mathscore
> head(Y)
     school mathscore
[1,]      1     52.11
[2,]      1     57.65
[3,]      1     66.44
[4,]      1     44.68
[5,]      1     40.57
[6,]      1     35.04
> ### weakly informative priors
> nu.0<-1
> sigmasq.0<-100
> eta.0<-1
> tausq.0<-100
> mu.0<-50
> gammasq.0<-25
> ###
>
> ### starting values
> m <- length(unique(Y[,1])) # number of schools
> n<-sv<-ybar<-rep(NA,m)
> for(j in 1:m)
+ {
+   ybar[j]<-mean(Y[Y[,1]==j,2])
+   sv[j]<-var(Y[Y[,1]==j,2])
+   n[j]<-sum(Y[,1] ==j)
+ }
> theta<-ybar
> sigma2<-mean(sv)
> mu<-mean(theta)
> tau2<-var(theta)
> ###
>
> ### setup MCMC
> set.seed(1)
> S<-5000
> THETA<-matrix( nrow=S,ncol=m)
> MST<-matrix( nrow=S,ncol=3)
> ###
>
> ### MCMC algorithm
> for(s in 1:S)
+ {
+
+   # sample new values of the thetas
+   for(j in 1:m)
+   {
```

```
+     vtheta<-1/(n[j]/sigma2+1/tau2)
+     etheta<-vtheta*(ybar[j]*n[j]/sigma2+mu/tau2)
+     theta[j]<-rnorm(1,etheta,sqrt(vtheta))
+   }
+
+   #sample new value of sigma2
+   nun<-nu0+sum(n)
+   ss<-nu0*sigmasq.0;
+   for(j in 1:m){
+     ss<-ss+sum((Y[[j]]-theta[j])^2)
+   }
+   sigma2<-1/rgamma(1,nun/2,ss/2)
+
+   #sample a new value of mu
+   vmu<- 1/(m/tau2+1/gammasq.0)
+   emu<- vmu*(m*mean(theta)/tau2 + mu.0/gammasq.0)
+   mu<-rnorm(1,emu,sqrt(vmu))
+
+   # sample a new value of tau2
+   etam<-eta.0+m
+   ss<- eta.0*tausq.0 + sum( (theta-mu)^2 )
+   tau2<-1/rgamma(1,etam/2,ss/2)
+
+   #store results
+   THETA[s,]<-theta
+   MST[s,]<-c(mu,sigma2,tau2)
+ }
```

Now consider the following plot that contains the posterior distribution for school 46 and school 48, along with the data points for each plotted along the bottom. Note the large circle represents the sample mean for each school. Comment on the differences between the sample means and the means of the posterior distributions. Why does this happen and is it a good thing?

Shrinkage

Recall the posterior mean can be represented as a weighted average, specifically in this case:

$$E[\theta_j|\tilde{y}_j, \mu, \tau^2, \sigma^2] = \frac{\bar{y}_j n_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}. \tag{16}$$

In this case μ and τ^2 are not chosen parameters from prior distributions, but rather they come from the between group model. So the posterior means for test scores at each school are pulled from the sample mean toward the overall group mean across all of the schools. This phenemenon is known as **shrinkage**.

Schools with more students taking the exam see less shrinkage, as there is more weight on the data given more observations. So the figure we discussed before, shows more shrinkage for school 82 as there were fewer observations.

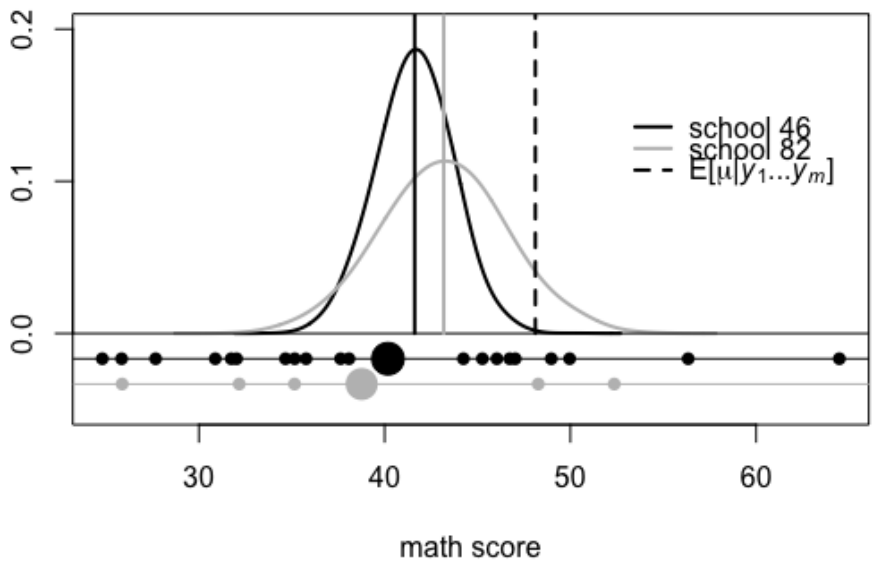


Figure 6: Posterior distributions and data points for schools 46 and 82.

So what about shrinkage, does it make sense? Is it a good thing?

We will soon see that it is an extremely powerful tool and actually dominates the unbiased estimator (the MLE for each distribution). This suprising result is commonly known as Stein's Paradox.

Hierarchical Modeling of Means and Variances

The model we just described and fit was somewhat restrictive in that each school was known to have a common variance. It is likely that schools with a more heterogenous mix of students would have greater variance in the test scores. There are a couple of solutions, the first involves a set of i.i.d. priors on each σ_j^2

$$\sigma_1^2, \ldots, \sigma_m^2 \sim \text{i.i.d. } gamma(\nu_0/2, \nu_0\sigma_0^2/2), \tag{17}$$

however, this results in a full conditional distribution for σ_j that only takes advantage of data from school j. In other words no information from the other schools is used to estimate that variance.

Another option is to consider ν_0 and σ_0^2 as parameters to be estimated in the hierarchical model. A com-

mon prior for σ_0^2 would be $p(\sigma_0^2) \sim Gamma(a, b)$. Unfortunately, there is not a semi-conjugate prior distribution for ν_0. The textbook suggests a geometric distribution, where $p(\nu_0) \propto \exp(-\alpha\nu_0)$. Then the full conditional distribution allows a sampling procedure that enumerates over the domain of possible values. This procedure allows shrinkage for the variance terms as well. It is worth noting, that pooling variances is a common way to tackle this particular problem.

Week 10 Oct 31 - Nov 5

Bayesian Testing

Up until now, we've primarily concerned ourselves with estimation type problems. However, many perform hypothesis tests.

Say, $x \sim N(0, 1)$ there are three types of tests you might consider for testing μ:

1. $\mu < \mu_0$ (one-sided test)
2. $\mu = \mu_0$ (point test)
3. $\mu \in [\mu_1, \mu_2]$ (interval test)

In a Bayesian framework, we will use point mass priors for Bayesian hypothesis testing.

Example. Consider testing the hypothesis $H_0 : \theta = \theta_0$ vs $H_1 : \theta_0 \neq \theta_1$. Say you observe data, $\tilde{x} = (x_1, \ldots, x_n)$, where $x_i \sim N(\theta, \sigma^2)$ with σ^2 known.

- **Q:** How would this question be addressed in a classical framework?

- If we want to be Bayesian, we need a prior. Suppose we choose a flat prior, $p(\theta) \propto 1$. Then $p(\theta = \theta_0|\tilde{x}) \sim N(\bar{x}, \sigma^2/n)$. With this distribution, we compute $Pr(H_0|\tilde{x}) = 0$. This is a result of a continuous prior on a continuous parameter.

- Consider a different prior that places mass on $H_0 : \theta = \theta_0$ which is non-zero. Specifically let $Pr(\theta = \theta_0) = p_{\theta_0}(\theta) = \pi_{\theta_0} > 0$. Say $\pi_{\theta_0} = 0.5$.
- We also need a prior for the alternative space. Let's choose a conjugate prior. Let $\theta_1 \sim N(\mu, \tau^2)$.
- Combining these the prior is:

$$p(\theta) = \pi_0 \delta(\theta = \theta_0) + (1 - \pi_0) p_1(\theta), \tag{18}$$

  where $\delta(\theta = \theta_0)$ is an indicator function for $\theta = \theta_0$ and $p_1(\theta) = N(\mu_1, \tau^2)$. This prior is known as a point mass prior. It is also a special case of another prior known as a spike-and-slab prior.

- Recall $x_i \sim N(\theta, \sigma^2)$. We want to know $Pr(H_0|\tilde{x})$ and $Pr(H_1|\tilde{x})$.

$$\begin{aligned} Pr(H_0|\tilde{x}) &= \frac{p(\tilde{x}|H_0)p(H_0)}{p(\tilde{x})} \\ &\propto p(\tilde{x}|H_0)\pi_0 \\ &\propto \int_{\theta \in H_0} p(\tilde{x}|\theta_0)p(\theta_0)d\theta\pi_0 \end{aligned}$$

  and similarly,

$$Pr(H_1|\tilde{x}) \propto \int_{\theta \in H_1} p(\tilde{x}|\theta)p_1(\theta)d\theta(1 - \pi_0)$$

- So how do we pick $p_1(\theta)$?

  In this course we will use $p_1(\theta) \sim N(\mu_1, \tau^2)$. So how do we pick the parameters of this distribution μ and τ^2?

- This point mass mixture prior can be written as:

$$p(\theta) = \pi_0 \delta(\theta = \theta_0) + (1 - \pi_0) p_1(\theta) \tag{19}$$

- Consider the ratio:

$$\frac{p(\tilde{x}|H_0)}{p(\tilde{x}|H_1)} = \frac{\int_{\theta \in H_0} p(\tilde{x}|\theta_0)p_0(\theta)d\theta}{\int_{\theta \in H_1} p(\tilde{x}|\theta)p_1(\theta)d\theta} = \left(\frac{p(H_0|\tilde{x})}{p(H_1|\tilde{x})}\right) / \left(\frac{p(H_0)}{p(H_1)}\right) \tag{20}$$

  This is known as a Bayes Factor.

- Recall the maximum-likelihood has a related form:

$$\frac{max : \theta \in H_0 \;\; \mathcal{L}(\theta|\tilde{x})}{max : \theta \in H_1 \;\; \mathcal{L}(\theta|\tilde{x})} \tag{21}$$

  In a likelihood ratio test we compare the difference for specific values of θ that maximize the ratio, whereas the Bayes factor (BF) integrates out the parameter values - in effect averages across the parameter space.

- In this example, let's choose $\mu_1 = \theta_1$ and set $\tau^2 = \psi^2$. Note $\bar{x}$ is a sufficient statistic, so we consider $p(\bar{x}|\theta)$.Then:

$$\begin{aligned}
BF &= \frac{\int_{\theta \in H_0} p(\bar{x}|\theta) p_{\theta_0}(\theta) d\theta}{\int_{\theta \in H_1} p(\bar{x}|\theta) p_1(\theta) d\theta} = \frac{\sqrt{n}/\sigma \exp\left(-\frac{(\bar{x}-\theta_0)^2}{2\sigma^2 n}\right)}{1/\sqrt{\sigma^2/n + \psi^2} \exp\left(-\frac{(\bar{x}-\theta_0)^2}{2(\sigma^2/n+\psi^2}\right)} \\
&= \left(\frac{\sigma^2/n}{\sigma^2/n + \psi^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\left[\frac{(\bar{x}-\theta_0)^2}{\sigma^2/n} - \frac{(\bar{x}-\theta_0)^2}{\sigma^2/n + \psi^2}\right]\right) \\
&= \left(1 + \frac{\psi^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2}\left[\frac{(\bar{x}-\theta_0)^2}{\sigma^2/n}\left(1 + \frac{\sigma^2}{n\psi^2}\right)^{-1}\right]\right) \\
&= \left(1 + \frac{\psi^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2}\left[z^2\left(1 + \frac{\sigma^2}{n\psi^2}\right)^{-1}\right]\right).
\end{aligned}$$

  Note that $Pr(H_0|Data) = \left(1 + \frac{1-\pi_0}{\pi_0} BF^{-1}\right)$, (HW problem).

- Example. Let π_0 = 1/2, $\sigma^2 = \psi^2$, $N = 15$, $Z = 1.96$, plugging this all in we get BF = 0.66. This implies:

$$Pr(H_0|\bar{x}) = (1 + .66^{-1})^{-1} = 0.4.$$

  Q: Reject or not?

  Q: What is the corresponding p-value here?

  Consider the following scenarios with $z = 1.96$.

N	5	10	50	100	1000
$Pr(H_0\|\bar{x})$	.331	.367	.521	.600	.823

  In each case the p-value is 0.05. Note that for a given effect size (ψ^2) the Bayes Factor is effect size calibrated. For a given effect size, a p-value goes to zero. Hence the disagreement between

'practical significance' and 'statistical significance'.

- So in this case the relevant question is how to choose ψ^2. ψ is the distance we find meaningful for rejecting H_0. That is, if inferences about θ tend (with high probability) to be larger than ψ from θ_0 then reject.

- **Q:** What happens as $\psi^2 \to \infty$?

  Recall from our example:

$$BF = \left(1 + \frac{N\psi^2}{\sigma^2}\right)^{1/2} \exp\left(-1/2z^2\left[1 + \frac{\sigma^2}{n\psi^2}\right]^{-1}\right)$$

  so the BF $\to \infty$. This implies that we need to put proper priors on the parameters when using BF.

- Consider two models: $M_1 \& M_2$, each with parameters sets $\Theta^{(M_1)}$ and $\Theta^{(M_2)}$. The Bayes Factor is:

$$BF = \frac{\int \mathcal{L}(\Theta^{(M_1)}|\tilde{x}) p_{M_1}(\theta^{(M_1)}) d\Theta^{(M_1)}}{\int \mathcal{L}(\Theta^{(M_2)}|\tilde{x}) p_{M_2}(\theta^{(M_2)}) d\Theta^{(M_2)}} \tag{22}$$

  When $p_{M_1}(\Theta^{(M_1)}$ and $p_{M_2}(\theta^{(M_2)}$ are proper, the BF is well defined.

- **Q:** Can you ever specify an improper prior on any of the parameters in $\Theta^{(M_1)}$ and $\Theta^{(M_2)}$?

  Yes, so long as the parameter appears in both models (and that parameter has the same meaning in both models). Otherwise, the BF is meaningless if the parameter does not appear in both models.

Bayesian Regression

Linear modeling is an important element in a statistician's toolbox. We are going to discuss the impact of different priors versus the classical regression setting.

A common challenge in regression framework is variable selection (or model selection).
Q: How do you currently handle variable selection?

The Bayesian paradigm, through placing priors on the model space, provide a natural way to carryout model selection as well as model averaging.

Notation

In a regression framework the goal is to model the relationship between a variable of interest, y, and a set of covariates $\mathcal{X}$. Specifically, we are modeling the conditional expectation for y given a set of parameters $\mathcal{X}$, which can be formulated as:

$$E[y|\mathcal{X}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \tilde{\beta}^T \tilde{x}. \tag{23}$$

While this model is linear in the parameters, transformation of the covariates and basis functions give a great deal of flexibility and make linear regression a powerful tool.

Typically, the model is stated as:

$$y_i = \tilde{\beta}^T \tilde{x}_i + \epsilon_i \tag{24}$$

where the $\epsilon_i' s$ are i.i.d. from a $N(0, \sigma^2)$ distribution. Recall, we could also think about regression with error terms from the $t-$ distribution as well.

Using the normal distributional assumptions then the joint distribution of the observed data, given the data $x_1, \ldots, x_n$ along with β and σ^2 can be written as:

$$\begin{aligned} p(y_1, \ldots, y_n | \tilde{x}_1, \ldots, \tilde{x_n}, \tilde{\beta}, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \tilde{x}_i, \tilde{\beta}, \sigma^2) & (25) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}^T \tilde{x}_i \right)^2 \right]. & (26) \end{aligned}$$

Note this is the same as the sampling, or generative model, that we have seen earlier in class.

Given our newfound excellence in linear algebra, the model is often formulated using matrix expressions and a multivariate normal distribution. Let

$$\tilde{y} | X, \tilde{\beta}, \sigma^2 \sim MVN(X\tilde{\beta}, \sigma^2 I), \tag{27}$$

where $\tilde{y}$ is an $n \times 1$ vector of the responses, X is an $n \times p$ matrix of the covariates where the i^{th} row is $\tilde{x}_i$, and I is a $p \times p$ identity matrix.

In a classical setting, typically least squares methods are used tom compute the values of the covariates in a regression setting. Note in a normal setting these correspond to maximum likelihood estimates. Specifically, we seek to minimize the sum of squared residuals (SSR), where $SSR(\tilde{\beta}) = \left(\tilde{y} - X\tilde{\beta} \right)^T \left(\tilde{y} - X\tilde{\beta} \right)$.

Thus we will take the derivative of this function with respect to β to minimize this expression.

$$\begin{aligned}
\frac{d}{d\tilde{\beta}} SSR(\tilde{\beta}) &= \frac{d}{d\tilde{\beta}} \left(\tilde{y} - X\tilde{\beta}\right)^T \left(\tilde{y} - X\tilde{\beta}\right) && (28)\\
&= \frac{d}{d\tilde{\beta}} \left(\tilde{y}^T\tilde{y} - 2\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T X^T X\tilde{\beta}\right) && (29)\\
&= -2X^T\tilde{y} + 2X^T X\tilde{\beta} && (30)\\
\text{then set = 0} \quad \text{which implies} \quad & X^T X\beta = X^t\tilde{y} && (31)\\
\text{and} \quad & \tilde{\beta} = (X^T X)^{-1} X^t\tilde{y}. && (32)
\end{aligned}$$

This value is the OLS estimate of $\tilde{\beta}_{OLS} = (X^T X)^{-1} X^T\tilde{y}$. Under the flat prior $p(\tilde{\beta}) \propto 1$, $\tilde{\beta}_{OLS}$ is the mean of the posterior distribution.

Bayesian Regression

As we have seen, the sampling distribution is:

$$\begin{aligned}
p(\tilde{y}|X, \tilde{\beta}, \sigma^2) &\propto \exp\left[-\frac{1}{2\sigma^2}(\tilde{y} - X\tilde{\beta})^T(\tilde{y} - X\tilde{\beta})\right] && (33)\\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(\tilde{y}^T\tilde{y} - 2\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T X^T X\tilde{\beta}\right)\right] && (34)
\end{aligned}$$

Given this looks like the kernel of a normal distribution for $\tilde{\beta}$, we will consider a prior for $\tilde{\beta}$ from the normal family.

Let $\tilde{\beta} \sim MVN(\tilde{\beta}_0, \Sigma_0)$, then

$$\begin{aligned}
p(\tilde{\beta}|\tilde{y}, X, \sigma^2) &\propto p(\tilde{y}|X, \tilde{\beta}, \sigma^2) \times p(\tilde{\beta}) && (35)\\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(\tilde{\beta}^T X^T X\tilde{\beta} - 2\tilde{\beta}^T X^T\tilde{y}\right) - \frac{1}{2}\left(\tilde{\beta}^T\Sigma_0^{-1}\tilde{\beta} - 2\tilde{\beta}\Sigma_0^{-1}\tilde{\beta}_0\right)\right] && (36)\\
&\propto \exp\left[-\frac{1}{2}\left(\tilde{\beta}^T\left(\Sigma_0^{-1} + X^T X/\sigma^2\right)\tilde{\beta} - 2\tilde{\beta}\left(\Sigma_0^{-1}\tilde{\beta}_0 + X^t\tilde{y}/\sigma^2\right)\right)\right] && (37)
\end{aligned}$$

Thus, using the properties of the multivariate normal distribution from a previous chapter, we can identify the mean and variance of the posterior distribution.

- $Var(\tilde{\beta}|\tilde{y}, X, \sigma^2) = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1}$.
- $E[\tilde{\beta}|\tilde{y}, X, \sigma^2] = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1} \times \left(\Sigma_0^{-1}\tilde{\beta}_0 + X^T\tilde{y}/\sigma^2\right)$

For a sanity check, let's look at the posterior distribution under a flat prior, $p(\tilde{\beta}) \propto 1$. Then $p(\tilde{\beta}|-) \sim N\left((X^TX)^{-1}X^t\tilde{y}, (X^TX)^{-1}\sigma^2\right)$. Note these are the OLS estimates.

We still need to consider a prior on σ^2. As we have seen in other scenarios the semi-conjugate prior is from the Inverse Gamma distribution. Let $\sigma^2 \sim IG(\nu_0/2, \nu_0\sigma_0^2/2)$ then

$$\begin{aligned} p(\sigma^2|-) &\propto p(\tilde{y}|X, \tilde{\beta}, \sigma^2)p(\sigma^2) && (38)\\ &\propto \left[(\sigma^2)^{-n/2}\exp(-SSR(\tilde{\beta})/2\sigma^2)\right] \times \left[(\sigma^2)^{-\nu_0/2-1}\exp(-\nu_0\sigma_0^2/2\sigma^2)\right] && (39)\\ &\propto \left(\sigma^2\right)^{-\frac{\nu_0+n}{2}-1}\exp\left(-\left[SSR(\tilde{\beta}) + \nu_0\sigma_0^2\right]/2\sigma^2\right) && (40)\end{aligned}$$

We recognize this distribution as an

$$IG\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2 + SSR(\tilde{\beta})}{2}\right)$$

.

Given the full conditional distributions, we need to sketch out a Gibbs sampler to take draws from the full conditional distributions.

1. Update $\tilde{\beta}$:
   (a) Compute $V = Var(\tilde{\beta}|\tilde{y}, X, \sigma^2)$ and $E = E[\tilde{\beta}|\tilde{y}, X, \sigma^{2(s)}]$.
   (b) Sample $\tilde{\beta}^{(s+1)} \sim MVN(E, V)$
2. Update σ^2:
   (a) Compute $SSR(\tilde{\beta}^{(s+1)})$
   (b) Sample $\sigma^{2(s+1)} \sim IG\left([\nu_0+n]/2, [\nu_0\sigma_0^2 + SSR(\tilde{\beta}^{(s+1)})]/2\right)$.

Priors for Bayesian Regression

For this model we need to think about priors for σ^2 and $\tilde{\beta}$. From similar situations, we have a good handle on how to think about the prior on σ^2. In particular, the Inverse-Gamma distribution gives us a semi-conjugate prior and the parameterization using ν_0 and σ_0^2 provides and intuitive way to think about the parameters in this distribution. Recall, $p\tilde{\beta} = N(\tilde{\beta}_0, \Sigma_0)$. The challenge is how to come up for values for $\tilde{\beta}_0$ and Σ_0.

In an applied setting, often some information is available about the potential magnitude of the covariates.

This allows reasonable values for $\tilde{\beta}_0$ and the variance components in Σ_0, but still requires some thought for the covariance terms in Σ_0. As p, the number of covariates, increases this becomes more and more difficult.

The textbook discusses the *unit information prior*, that injects information proportional to a single observation - similar to how we have used ν_0 and η_0 previously in the course. One popular prior from this principle is known as Zellner's g-prior, where $\Sigma_0 = g\sigma^2(X^TX)^{-1}$. Using Zellner's g-prior the marginal distribution $p(\sigma^2|\tilde{y}, X)$ can be derived directly. This allows the conditional draws from $p(\tilde{\beta}|\sigma^2, \tilde{y}, X)$ in a similar fashion to our first normal settings when $p(\beta) = N(\mu_0, \sigma^2/\kappa_0)$.

Other common strategies are to use a weakly informative prior on Σ_0, say $\Sigma_0 = \tau_0^2 \times I_p$.

Bayesian Modeling and Ridge Regression

Ordinary Least Squares (OLS) regression can be written as:

$$\hat{\tilde{\beta}}_{OLS} = \text{arg min}_{\hat{\tilde{\beta}}} \ ||\tilde{y} - X\hat{\tilde{\beta}}||_2 \rightarrow \hat{\tilde{\beta}} = (X^TX)^{-1}X^T\tilde{y}, \tag{41}$$

where $||\tilde{x}||_p = (|x_1|^p + \cdots + |x_m|^p)^{1/p}$ is an LP norm. So the L2 norm is $||\tilde{x}||_2 = \sqrt{x_1^2 + \ldots x_m^2}$.

Recall ridge regression is a form of penalized regression such that:

$$\hat{\tilde{\beta}}_R = \text{arg min}_{\hat{\tilde{\beta}}} \ ||\tilde{y} - X\hat{\tilde{\beta}}||_2 + \lambda||\hat{\tilde{\beta}}_R||_2 \rightarrow \hat{\tilde{\beta}}_R = \left(X^TX + \lambda I\right)^{-1} X^T\tilde{y}, \tag{42}$$

where λ is a tuning parameter that controls the amount of shrinkage. As λ gets large all of the values are shrunk toward 0. As λ goes to 0, the ridge regression estimator results in the OLS estimator. It can be shown that ridge regression results better predictive ability than OLS by reducing variance of the predicted values at the expense of bias. Note that typically the X values are assumed to be standardized, so that the intercept is not necessary

Q: How do we choose λ?

An alternative form of penalized regression is known as Least Absolute Shrinkage and Selection Operator (LASSO). The LASSO uses an L1 penalty such that:

$$\hat{\tilde{\beta}}_L = \text{arg min}_{\hat{\tilde{\beta}}} \ ||\tilde{y} - X\hat{\tilde{\beta}}||_2 + \lambda||\hat{\tilde{\beta}}_L||_1, \tag{43}$$

the L1 penalty results in $||\tilde{x}||_1 = |x_1| + \cdots + |x_m|$, which minimizes the absolute differences. The nice feature of LASSO, relative to ridge regression, is that coefficients are shrunk to 0 providing a way to do *variable selection.* One challenge with LASSO is coming up with proper distributional assumptions for inference about variables.

Consider the following prior $p(\tilde{\beta}) = N(0, I_p\tau^2)$. How does this relate to ridge regression? First compute the posterior distribution for $\tilde{\beta}$.

$$p(\tilde{\beta}|-) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\tilde{\beta}^T X^T X\tilde{\beta} - \frac{1}{\sigma^2}\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T\frac{I_p}{\tau^2}\tilde{\beta}\right)\right] \quad (44)$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\tilde{\beta}^T\left(X^T X + \frac{\sigma^2}{\tau^2}I_p\right)\tilde{\beta} - \frac{1}{\sigma^2}\tilde{\beta}^T X^T\tilde{y}\right)\right] \quad (45)$$

Thus $Var(\tilde{\beta}|-) = \left(X^T X + \frac{\sigma^2}{\tau^2}I_p\right)^{-1}\sigma^2$ and $E(\tilde{\beta}|-) = \left(X^T X + \frac{\sigma^2}{\tau^2}I_p\right)^{-1}X^t\tilde{y}$.
Q: does this look familiar?

Define: $\lambda = \frac{\sigma^2}{\tau^2}$. How about now?
This is essentially the ridge regression point estimate.

Note that in a similar fashion we can use a specific prior on $\tilde{\beta}$ to achieve LASSO properties. It is also important to clarify the differences between classical ridge regression (and Lasso) with the Bayesian analogs. In the Bayesian case we can still easily compute credible intervals to account for the uncertainty in our estimation. Interval calculations for inference are difficult in the these settings, particularly for Lasso.

Week 11 Nov 7 - Nov 11

Bayesian Model Selection

Recall, we discussed common model selection techniques:

- All subsets (use aic, bic) works for moderate p
- Backward selection
- Forward selection
- Backward - Forward Selection
- Cross Validation

However, in a Bayesian framework we have a coherent way to talk about model selection. Specifically, given a prior on the model space we can compute posterior probability for a given model.

In model selection for linear regression the goal is to decide which covariates to include in the model. To do this, we introduce a parameter $\tilde{z}$, where $z_i = 1$ if covariate i is in the model, otherwise $z_i = 0$. Then define $\beta_i = z_i \times b_i$. Note the b_i' are the real-values regression coefficients. For now we will ignore the intercept (standardizing the covariates). The regression equation now becomes:

$$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i. \tag{46}$$

Again thinking about the regression model as a conditional expectation, then for $p = 3$:

$$\begin{aligned} E[y|\tilde{x}, \tilde{b}, \tilde{z} = (1,0,1)] &= b_1 x_1 + b_3 x_3 \\ E[y|\tilde{x}, \tilde{b}, \tilde{z} = (0,1,0)] &= b_2 x_2. \end{aligned}$$

Note that the vector $\tilde{z}_a$ defines a model and is interchangable with the notation M_a. Now the goal is a probabilistic statement about $\tilde{z}_a$, specifically:

$$Pr(\tilde{z}_a|\tilde{y}, X) = \frac{p(\tilde{y}|X, \tilde{z}_a)p(\tilde{z}_a)}{\int_{z^*} p(\tilde{y}|X, \tilde{z}^*)p(\tilde{z}^*)d\tilde{z}^*}, \tag{47}$$

where X is a matrix of observed covariates. Of course, this requires a prior on z_a, which we will see momentarily.

For model comparison between model a and model b, an alternative way to express this is through the following (familiar) ratio:

$$BF(a,b) = \frac{p(\tilde{y}|X, \tilde{z}_a)}{p(\tilde{y}|X, \tilde{z}_b)} = \left(\frac{Pr(\tilde{z}_a|\tilde{y}, X)}{Pr(\tilde{z}_b|\tilde{y}, X)}\right) / \left(\frac{p(\tilde{z}_a)}{p(\tilde{z}_b)}\right). \tag{48}$$

Of course this is a Bayes Factor.

Now the question is, how do we think about the priors for $\tilde{z}$ or equivalently for M_a?

The textbook does not explicitly discuss this, but a couple common parameters are the discrete uniform prior, where each model has the same prior probability. In terms of the $z_i's$ this would be equivalent to prior inclusion probability of 0.5. In other situations, a prior can be placed on the total number of parameters in the model.

Bayesian Model Comparison

The posterior probability for model a is a function of the prior $p(z_a)$ and $p(\tilde{y}|X, \tilde{z}_a)$ which is known as the marginal probability. In a regression setting the marginal probability is computed by integrating out the parameters as:

$$\begin{aligned} p(\tilde{y}|X, \tilde{z}_a) &= \int\int p(\tilde{y}, \tilde{\beta}_a, \sigma^2|X, \tilde{z}_a) d\tilde{\beta}_a d\sigma^2 \quad (49) \\ &= \int\int p(\tilde{y}|\tilde{\beta}_a, X) p(\tilde{\beta}_a|\tilde{z}_a, \sigma^2) p(\sigma^2) d\tilde{\beta}_a d\sigma^2, \quad (50) \end{aligned}$$

where $\tilde{\beta}_a$ is a $p_{z_a} \times 1$ vector containing the p_{z_a} elements in M_a. In general this integration is very difficult, particularly when p, the dimension of $\tilde{\beta}$, is large. However, recall Zellner's g-prior had a form that facilitates efficient integration. Under this prior $p(\tilde{\beta}_a|\sigma^2, \tilde{z}_a) = MVN_{p_{z_a}}\left(\tilde{\beta}_0, g\sigma^2(X^TX)^{-1}\right)$.

It can be shown that integrating out $\tilde{\beta}_a$, $\int p(\tilde{y}|X, \tilde{z}_a, \sigma^2) = p(\tilde{y}|\tilde{\beta}_a, X) p(\tilde{\beta}_a|\tilde{z}_a, \sigma^2) d\tilde{\beta}_a$ is fairly straightforward. This leaves:

$$p(\tilde{y}|X, \tilde{z}_a) = \int p(\tilde{y}|X, \tilde{z}_a, \sigma^2) p(\sigma^2) d\sigma^2. \quad (51)$$

Due to the form of the priors, this can also be easily integrated such that the marginal probability is:

$$p(\tilde{y}|X, \tilde{z}_a) = \pi^{-n/2} \frac{\Gamma([\nu_0 + n]/2}{\Gamma(\nu_0/2} (1+g)^{-p_{z_a}/2} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + SSR_g^z)^{(\nu_0+n)/2}}, \quad (52)$$

where $SSR_g^z = \tilde{y}^T(I_{p_z} - \frac{g}{g+1}X(X^TX)^{-1}X^T)\tilde{y}$.

Given the marginal likelihoods, we can compute the posterior probability of a given model, M_a as:

$$Pr(M_a = \tilde{y}, X) = Pr(\tilde{z}_a|\tilde{y}, X) = \frac{p(\tilde{y}|X, \tilde{z}_a) p(\tilde{z}_a)}{\int_{z^*} p(\tilde{y}|X, \tilde{z}^*) p(\tilde{z}^*) d\tilde{z}^*}. \quad (53)$$

Using this formulation we can choose the most probable model or a set of the most probable models.

Bayesian Model Averaging

A powerful tool in Bayesian modeling, particularly for predictive settings, is Bayesian model averaging. Rather than choosing a single model, or set of models, we will average across models according to their posterior probability.

Assume we are interested in some quantity of interest Δ, which can be computed as a function of the posterior distribution. Then:

$$p(\Delta|X,\tilde{y}) = \sum_{i=1}^{k} p(\Delta|M_k, X, \tilde{y}) Pr(M_k|X,\tilde{y}) \tag{54}$$

.

For example let Δ represent the posterior predictive distribution for $\tilde{y}^*$, given X^*. Then the model averaged posterior predictive distribution can be written as,

$$p(\tilde{y}^*|, X^* X, \tilde{y}) = \sum_{i=1}^{k} p(\tilde{y}^*|X^*, M_k, X, \tilde{y}) Pr(M_k|X,\tilde{y}). \tag{55}$$

This model averaged prediction is a special type of an ensemble method, which have nice predictive properties *and* in this case account for uncertainty in the model selection process.

General Model Selection and Averaging

In many cases, the g-prior framework is too restrictive or the models will not allow closed form solutions for the marginal likelihood. For instance consider the following model:

$$\tilde{y}|- \sim N(XB, \sigma^2 H(\phi) + \tau^2 I), \tag{56}$$

where $H(\phi)$ is a correlation matrix. Finding the marginal (integrated) likelihood analytically would be very difficult in this case. It turns out this model is often used in Bayesian spatial modeling. By introducing an infinite-dimensional Gaussian distribution known as a Gaussian Process (GP), a posterior predictive distribution can be computed for any point in space. This is a Bayesian analog to Kriging.

In situations like this, where model selection is often conducted using MCMC. Micaela will give us an overview of Gibbs sampling for linear regression, but the basic idea is that each iteration you:

- Update each z_i
- Sample σ^2
- Sample $\tilde{\beta}|\tilde{z}$,

where for each z_i the $Pr(z_i = 1|\tilde{y}, X, \tilde{z}_{-i})$ can be computed and $\tilde{z}_{-i}$ is all of the elements excluding i.