A First Course in Bayesian Statistical Methods notes

Shua

Update: September 3, 2023

Contents

1 Introduction and examples

1.1 Introduction

Bayesian inference: the process of learning by updating prior probabilistic beliefs in light of new information. Data analysis tools built on these foundations are known as **Bayesian methods**.

1.1.1 Bayesian learning framework

The numerical values of both the population characteristics and the dataset are uncertain. After a dataset y is obtained, the information it contains can be used to decrease our uncertainty about the population characteristics. Quantifying this change in uncertainty is the purpose of Bayesian inference.

In fact, we want to estimate a parameter $\theta \in \Theta$ from a dataset $y \in \mathcal{Y}$.

- $\mathcal{Y}$: The *sample space*, the set of all possible datasets, from which a single dataset y will result.
- Θ: The *parameter space*, the set of possible parameter values, from which we identify the value that best represents the true characteristics.
- $p(\theta)$: The *prior distribution*, describes our belief that θ represents the true characteristics. (for each $\theta \in \Theta$)
- $P(y \mid \theta)$: The *sampling model*, describes the probability that of a specific dataset given a parameter. (for each $\theta \in \Theta$ and $y \in \mathcal{Y}$)

The *posterior distribution* is obtained from the prior distribution and sampling model via *Bayes' rule*:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_\Theta p(y \mid \tilde{\theta})p(\tilde{\theta})\, d\tilde{\theta}}.$$

Note that the denominator is constant and doesn't need to be computed, since we can just normalize our posterior distribution such that $P(\theta \mid y)$ for all Θ sums up to 1. Thus we commonly write

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta).$$

1.1.2 Versus frequentist learning

The difference between Bayesian learning and frequentist learning is the consideration of *prior beliefs* about parameters. In standard Maximum Likelihood Estimation (MLE), we select the parameter that is most likely to have generated the observed data:

$$\theta_{ML} = \operatorname*{argmax}_\theta p(y \mid \theta).$$

Using Bayesian Maximum A Posteriori Estimation, we select θ that is most likely given the observed data. The difference is that our measure of "likelihood given the data" is influenced by prior belief about θ:

$$\theta_{MAP} = \operatorname*{argmax}_\theta p(\theta \mid y) = \operatorname*{argmax}_\theta p(y \mid \theta)p(\theta).$$

Note that with an uninformative prior $\theta \sim$ Uniform, the MAP estimate is the same as the ML estimate.

1.2 Why Bayes?

1. If $p(\theta)$ approximates our beliefs, then the fact that $p(\theta \mid y)$ is optimal under $p(\theta)$ means that it will also generally serve as a good approximation to what our posterior beliefs should be.
2. We may want to use Bayes' rule to explore how the data would update the beliefs of a variety of people with differing prior opinions.
3. In many complicated statistical problems there are no obvious non-Bayesian methods of estimation or inference. In these situations, Bayes' rule can be used to generate estimation procedures, and the performance of these procedures can be evaluated using non-Bayesian criteria.
4. As an approach to probability and statistics:
   Bayesian methods have excellent practical benefits as data analysis tools:
   (a). Even if prior probabilities are not exactly quantifiable, approximations of $p(\theta)$ and $p(\theta \mid y)$ are still useful for analyzing how rational learners would change beliefs
   (b). Bayesian methods can represent principled ways of doing analysis when there are no alternative methods
5. As models of cognition:
   An appeal of Bayesian learning is that it is also cognitively intuitive. Humans have beliefs about the world, whose uncertainty can be expressed probabilistically. Then, given data, these beliefs are rationally updated.
   Bayesianism is not without its detractors, however. Some critics argue that the evidence that Bayesian analysis is weak, and that sufficiently sophisticated models are unfalsifiable. See [Bowers & Davis (2012)], comment by [Griffiths, Chater, Norris, Pouget (2012)], reply by [Bowers & Davis (2012)].

1.2.1 Sensitivity analysis

If we change belief in prior, we get different posterior distributions. The more "peaked" prior is, the less peaked the posterior will be given a $Y = 0$ (and the less the Bayesian solution will approximate the ML estimate).

Quantify how changes in the prior beliefs affect our posterior estimates:

Recall the expectation and variance of Beta distributions. If $\theta \sim \text{Beta}(\alpha, \beta)$, then

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Due to the properties of these functions, we can parameterize the Beta distribution alternatively with Expectation: $\theta_0 = \frac{\alpha}{\alpha+\beta}$ and Precision: $w = a + b$.

Since $(\theta \mid Y = y) \sim \text{Beta}(a + y, b + n - y)$,

$$\mathbb{E}(\theta \mid Y = y) = \frac{a + y}{a + b + n} = \frac{n}{a + b + n}\frac{y}{n} + \frac{a + b}{a + b + n}\frac{a}{a + b} = \frac{n}{w + n}\bar{y} + \frac{w}{w + n}\theta_0.$$

The posterior expectation is a weighted average of the sample mean $\bar{y}$ and the prior expectation θ_0. In terms of estimating θ, θ_0 represents our prior guess at the true value of θ and w represents our confidence in this guess, expressed on the same scale as the sample size.

We can compute such a posterior distribution for a wide range of θ_0 and w values to perform a sensitivity analysis, an exploration of how posterior information is affected by differences in prior opinion.

1.2.2 Comparison to non-Bayesian methods

When we use the frequentist maximum likelihood estimator, we get an estimated $\theta_{ML} = 0$. Since our estimate is subject to sampling error, we commonly construct confidence intervals for these estimates.

The *Wald interval* is a commonly used confidene interval for a population proportion. However, it is not meant to be used for small sample sizes or situations in which the observed proportion is close to (or equals) 0 or 1, since in these cases the error of a binomially-distributed observation is not at all like the normal distribution For an observation $Y = 20$, for example, the Wald CI is, regardless of the level of confidence, just 0. We wouldn't want to say with 99.999% confidence that the population mean is 0, given our small sample size.

The previous Bayesian estimate, however, works well for both small and large n. With small n, the estimator allows us to encode prior beliefs about the true proportion. With w and θ_0 as before:

$$\hat{\theta} = \mathbb{E}(\theta \mid Y = y) = \frac{n}{n+w}\frac{y}{n} + \frac{w}{n+w}\theta_0.$$

Notice that this is kind of an average between the prior expectation θ_0 and the observed proportion of the data $\frac{y}{n}$, weighted by the amount of data n. For large n, $\hat{\theta}$ becomes dominated by the data, regardless of prior estimate and confidence.

Theoretical details on the properties of Bayesian estimators are covered later in Section 5.4.

1.2.3 Building a predictive model

A brief synopsis of an example is in Chapter 9, where we want to build a predictive model of diabetes progression from 64 variables such as age, sex and BMI.

We will first estimate the parameters in a regression model using a "training" dataset consisting of measurements from 342 patients. We will then evaluate the predictive performance of the estimated regression model using a separate "test" dataset of 100 patients.

- **Sampling model and parameter space**
  Consider linear regression models of the form

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{64} x_{i,64} + \sigma\epsilon_i$$

- **Prior distribution**
  Defining a joint prior probability distribution for 65 parameters $\Rightarrow$ near-impossible task.
  Use a prior distribution that only represents some aspects of our prior beliefs.
- Posterior distribution
  Given data $\boldsymbol{y} = (y_1, \cdots, y_{342})$ and $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{342})$, the posterior distribution $p(\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbf{X})$ can be computed and used to obtain $\Pr(\beta_j \neq 0 \mid \boldsymbol{y}, \mathbf{X})$for each regression coefficient j.
- **Predictive performance and comparison to non-Bayesian methods**
  We can evaluate how well this model performs by using it to predict the test data:
  Let $\hat{\boldsymbol{\beta}}_{Bayes} = E[\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbf{X}]$ be the posterior expectation of $\boldsymbol{\beta}$, and let $\mathbf{X}_{test}$ be the test dataset.
  Use $\hat{\boldsymbol{y}}_{test} = \mathbf{X}\hat{\boldsymbol{\beta}}_{Bayes}$ to compute a predicted value for each of the 100 observations.
  The non-Bayesian approach (most commonly: the ordinary least squares (OLS) estimate): the value $\hat{\boldsymbol{\beta}}_{\text{ols}}$

of $\boldsymbol{\beta}$ that minimizes the sum of squares of the residuals (SSR) for the observed data:

$$\mathrm{SSR}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2,$$

and is given by the formula $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$.
Predictions for the test data based on this estimate are given by $\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}}$.
Thus, Compare $\sum (y_{\text{test},i} - \hat{y}_{\text{test},i})^2/100$

- **Result**

Bayesian regression does better than standard linear regression. The standard ordinary least squares (OLS) estimate of β does worse than the Bayesian method on the test set.
This is due to overfitting, and OLS's "inability to recognize when the sample size is too small to accurately estimate the regression coefficients."
The standard remedy to this problem is to fit a "sparse" regression model, in which some or many of the regression coefficients are set to zero.

1. One method of choosing which coefficients to set to zero is the Bayesian approach described above.
2. Another popular method is the "lasso". The lasso estimate is the value $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ of $\boldsymbol{\beta}$ that minimizes $\mathrm{SSR}(\boldsymbol{\beta} : \lambda)$, a modified version of the sum of squared residuals:

$$\mathrm{SSR}(\boldsymbol{\beta} : \lambda) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,.$$

2 Belief, probability and exchangeability

1. Discuss what properties a reasonable belief function should have, show probabilities have these properties.
2. Review the basic machinery of discrete and continuous random variables and probability distributions.
3. Explore the link between independence and exchangeability.

2.1 Belief functions and probabilities

2.1.1 Belief functions

Let F, G, and H be events. A *belief function* $\text{Be}(\cdot)$ should correspond to certain intuitions about our beliefs about the likelihood of events.

1. $\text{Be}(F) > \text{Be}(G)$ means we would prefer to bet F is true than G is true
2. $\text{Be}(F \mid H) > \text{Be}(G \mid H)$ means that if we knew that H were true, then we would prefer to bet that F is also true than bet G is also true.
3. $\text{Be}(F \mid G) > \text{Be}(F \mid H)$ means that if we were forced to bet on F, we would prefer to do it under the condition that G is true rather than H is true

2.1.2 Axioms of beliefs

Any function that is to numerically represent our beliefs may have the following properties:

1. (**B1**) $\text{Be}(\text{not } H \mid H) \leq \text{Be}(F \mid H) \leq \text{Be}(H \mid H)$

2. (**B2**) $\text{Be}(F \text{ or } G \mid H) \geq \max\{\text{Be}(F \mid H), \text{Be}(G \mid H)\}$
3. (**B3**) $\text{Be}(F \text{ and } G \mid H)$ can be derived from $\text{Be}(G \mid H)$ and $\text{Be}(F \mid G \text{ and } H)$

2.1.3 Axioms of probability

Probability functions have axioms that satisfy our notions of belief:

1. (**P1**) Contradictions and tautologies: $0 = P(\text{not } H \mid H) \leq P(F \mid H) \leq P(H \mid H) = 1$
2. (**P2**) Addition rule: $P(F \cup G \mid H) = P(F \mid H) + P(G \mid H)$ if $F \cap G = \emptyset$
3. (**P3**) Multiplication rule: $P(F \cap G \mid H) = P(G \mid H)P(F \mid G \cap H)$

Note that the axioms of probability and theorems discussed in this section are the same whether you subscribe to a Bayesian or frequentist interpretation of probability.

2.2 Events, partitions and Bayes' rule

Consider a set $\mathcal{H}$, which is the "set of all possible truths." We can partition $\mathcal{H}$ into discrete subsets $\{H_1, \ldots, H_k\}$, where only one subset consists of the truth.

Definition 2.1 *(Partition) A collection of sets $\{H_1, \cdots, H_k\}$ is a partition of another set $\mathcal{H}$ if*

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

In the context of identifying which of several statements is true, if $\mathcal{H}$ is the set of all possible truths and $\{H_1, \cdots, H_k\}$ is a partition of $\mathcal{H}$, then exactly one out of $\{H_1, \cdots, H_k\}$ contains the truth.

2.2.1 Partitions and probability

We can assign probabilities whether each of these sets contains the truth. First, some event in $\mathcal{H}$ is true, so $P(\mathcal{H}) = 1$. Let E be some observation (in this case related to the truth of one H_i). Then,

- Rule of total probability: $\sum_i P(H_i) = 1$
- Marginal probability: $P(E) = \sum_i P(E \cap H_i) = \sum_i P(E \mid H_i)P(H_i)$
  - The total probability of an event occurring is the sum of all of its probabilities under the possible partitions of truths
- Bayes' rule: $P(H_i \mid E) = \frac{\overbrace{P(E \mid H_i)}^{\text{likelihood}}\overbrace{P(H_i)}^{\text{prior}}}{P(E)} = \frac{P(E \mid H_i)P(H_i)}{\sum_{k=1}^{K} P(E \mid H_k)P(H_k)}$

2.3 Independence

Definition 2.2 *(Independence)*

- *Two events F and G are independent if $P(F \cap G) = P(F)P(G)$.*
- *Two events F and G are conditionally independent given H if $P(F \cap G \mid H) = P(F \mid H)P(G \mid H)$.*

By Axiom **P3**, the following is always true: $P(F \cap G \mid H) = P(G \mid H)P(F \mid G \cap H)$. If F and G are conditionally independent given H, then we must have:

$$\Pr(G \mid H)\Pr(F \mid H \cap G) \overset{\text{always}}{=} \Pr(F \cap G \mid H) \overset{\text{independence}}{=} \Pr(F \mid H)\Pr(G \mid H)$$
$$\Pr(G \mid H)\Pr(F \mid H \cap G) = \Pr(F \mid H)\Pr(G \mid H)$$
$$\Pr(F \mid H \cap G) = \Pr(F \mid H).$$

T$P(F \mid H \cap G) = P(F \mid H)$ if we know about H, and F and G are conditionally independent given H, then knowing G does not change belief about H. This is a key property leveraged in Bayesian networks.

2.4 Random variables

In Bayesian inference, a random variable is defined as an unknown numerical quantity about which we make probability statements. Additionally, a fixed but unknown population parameter is also a random variable.

2.4.1 Discrete random variables

A random variable Y is *discrete* if the set of all its possible values $\mathcal{Y}$ is countable, i.e. they can be enumerated $\mathcal{Y} = \{y_1, y_2, \dots\}$.Examples include the binomial and Poisson distributions.

Discrete random variables have a *probability mass function* (PMF): $f(y) = P(Y = y)$ which assigns a certain probability to every discrete point in its sample space. From this probability mass function, a *cumulative distribution function* (CDF) is also defined:

$$F(y) = P(Y \leq y) = \sum_{y_i \leq y} f(y_i)$$

Or: The event that the outcome Y of our survey has the va lue y is expressed as $\{Y = y\}$. For each $y \in Y$, our shorthand notation for $\Pr(Y = y)$ will be $p(y)$.

This function of y is called the *probability density function* (pdf) of Y , and it has the following properties:

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$
3. General probability statements about Y can be derived from the pdf. For example, $\Pr(Y \in A) = \sum_{y \in A} p(y)$
4. If A and B are disjoint subsets of $\mathcal{Y}$, then

$$\Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) = \Pr(Y \in A) + \Pr(Y \in B) = \sum_{y \in A} p(y) + \sum_{y \in B} p(y).$$

2.4.2 Continuous random variables

Y is *continuous* if $\mathcal{Y}$ can take *any* value in an interval. Examples include the normal and beta distributions. The probability of Y taking a single value in the sample space is 0. So instead we describe such distributions with *probability density functions* (PDFs) $f(y)$, which must be integrated over an interval to obtain a probability:

$$P(a \leq y \leq b) = \int_a^b f(y)\, dy.$$

This function is called the probability density function of Y, and its properties are similar to those of a pdf for a discrete random variable: $0 \leq p(y)$ for all $y \in \mathcal{Y}$; $\int_{y \in \mathbb{R}} p(y)\, dy = 1$.

These variables also have cumulative distribution function (CDFs):

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(x)\, dx$$

$F(\infty) = 1$, $F(-\infty) = 0$, and $F(b) \leq F(a)$ if $b < a$. Probabilities of events can be derived from the cdf:

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

2.4.3 Descriptions of distributions

In the same way that we use the mean, mode, and median to describe samples, we can use them to describe distributions. Notice that for many distributions, these quantities are *not* the same.

Mean or expectation of an unknown quantity Y

$\mathrm{E}[Y] = \sum_{y \in \mathcal{Y}} yp(y)$ if Y is discrete; $\mathrm{E}[Y] = \int_{y \in \mathcal{Y}} yp(y)dy$ if Y is continuous.

Mode: the most probable value of Y

Median: the value of Y in the middle of the distribution."

We also use variance and quantiles to measure the *spread* of distributions.

Variance

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

2.5 Joint distributions

2.5.1 Discrete random variables

Let Y_1 and Y_2 be random variables with sample spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$.

The *joint pdf or joint density* of Y_1 and Y_2 is defined as:

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

The *marginal density* of Y_1 is obtained by summing over all possible values of Y_2:

$$p_{Y_1}(y_1) = p(y_1) \equiv \Pr(Y_1 = y_1) = \sum_{y_2 \in \mathcal{Y}_2} \Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) = \sum_{y_2 \in \mathcal{Y}_2} p(y_1, y_2) = \sum_{y_2 \in \mathcal{Y}_2} p(y_1 \mid y_2) p(y_2).$$

The *conditional density* of Y_2 given $\{Y_1 = y_1\}$ is

$$p_{Y_2|Y_1}(y_2 \mid y_1) = p(y_2 \mid y_1) = \frac{\Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{\Pr(Y_1 = y_1)} = \frac{p_{Y_1 Y_2}(y_1, y_2)}{p_{Y_1}(y_1)} = \frac{p(y_1, y_2)}{p(y_1)}.$$

Notice that given the joint density $p(y_1, y_2)$, we can calculate marginal and conditional densities $\{p(y_1), p(y_2), p(y_1 \mid y_2), p(y_2 \mid y_1)\}$ by simply summing up the relevant variables. Additionally, given $p(y_1)$ and $p(y_2 \mid y_1)$, (or the reverse), we can reconstruct the joint distribution. However, given only marginal densities $p(y_1)$ and $p(y_2)$, we can't reconstruct the joint distribution, since we don't know whether the events are independent.

2.5.2 Continuous random variables

In the continuous case, the probability density function is a function of y_1 and y_2 such that the CDF is

$$F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} p(y_1, y_2)\, dy_2\, dy_1.$$

Obtaining the marginal densities can be done by integrating out the irrelevant variable:

$$p(y_1) = \int_{-\infty}^{\infty} p(y_1, y_2)\, dy_2, \quad p(y_2) = \int_{-\infty}^{\infty} p(y_1, y_2)\, dy_1$$

With the marginal densities, you can compute the conditional densities

$$p(y_2 \mid y_1) = p(y_1, y_2)/p(y_1).$$

2.5.3 Bayes' rule and parameter estimation

$p(\theta)$, beliefs about θ; $p(y \mid \theta)$, beliefs about Y for each value of θ.

Having observed $\{Y = y\}$, we need to compute our updated beliefs about θ: $p(\theta \mid y) = p(\theta, y)/p(y) = p(\theta)p(y \mid \theta)/$
This conditional density is called the posterior density of θ.

As a function of θ, $p(\theta \mid y) = \dfrac{p(\theta)p(y \mid \theta)}{\int_\theta p(\theta)p(y \mid \theta)d\theta} \propto p(\theta)p(y \mid \theta).$

2.6 Independent random variables

Let $Y_1, \ldots, Y_n$ be random variables dependent on a common parameter θ. Then $Y_1 \ldots, Y_n$ are conditionally independent given θ if

$$p(y_1, \ldots, y_n \mid \theta) = p(y_1 \mid \theta) \times \cdots \times p(y_n \mid \theta).$$

Note this extends naturally from the definition of independent of two random variables, $P(A \cap B) = P(A)P(B)$. Thus, knowing about any Y_i does not give any information about the other Y_j. Lastly, the joint density of these variables can be defined as

$$p(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta).$$

We say that $Y_1, \ldots, Y_n$ are *conditionallly independent and identically distributed* (i.i.d.):

$$Y_1, \ldots, Y_n \mid \theta \sim \text{i.i.d. } p(y \mid \theta).$$

2.7 Exchangeability

In many situations with random variables, we would intuit that the specific order of observation of these random variables isn't important. Eg, consider a random sample of 3 participants from an infinite population which may or may not have a property (1 or 0). It makes sense that $p(0, 0, 1) = p(1, 0, 0) = p(0, 1, 0)$. since the likelihood of a person having the property or not is θ, regardless of the sample. This property is exchangeability.

Definition 2.3 *Exchangable. Let $Y_1, \ldots, Y_n$ be random variables. Let $p(y_1, \ldots, y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations π of $\{1, \ldots, n\}$, then $Y_1, \ldots, Y_n$ are exchangable.*

Claim: If $\theta \sim p(\theta)$ and $Y_1, \ldots Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \ldots Y_n$ are exchangeable.

Proof: Suppose $Y_1, \ldots, Y_n$ are conditionally i.i.d. given some unknown parameter θ.

Then for any permutation π of $\{1, \ldots, n\}$ and any set of values $(y_1, \ldots, y_n) \in \mathcal{Y}^n$,

$$\begin{aligned}
p(y_1, \ldots, y_n) &= \int p(y_1, \ldots, y_n \mid \theta) p(\theta)\, d\theta && \text{(definition of marginal probability)} \\
&= \int \left(\prod_{i=1}^{n} p(y_i \mid \theta) \right) p(\theta)\, d\theta && (Y_i\text{'s are conditionally i.i.d.}) \\
&= \int \left(\prod_{i=1}^{n} p(y_{\pi_i} \mid \theta) \right) p(\theta)\, d\theta && \text{(product does not depend on order)} \\
&= p(y_{\pi_1}, \ldots, y_{\pi_n}). && \text{(definition of marginal probability)}
\end{aligned}$$

Classical assumption of Bernoulli variables $X_1, X_2, \ldots, X_n$ as outcomes of the same experiment (e.g. a coin flip): *independence*. But continuing to observe X_js should result in a change of opinion about the distribution of coin flip outcomes (e.g. gradually learning coin bias). So Bayesian statisticians should assume *exchangeability*, a weaker condition than *independence*.

2.8 de Finetti's theorem

Theorem 2.1 *de Finetti's theorem. Let $Y_i \in \mathcal{Y}$ for all $i \in \{1, 2, \cdots\}$. Suppose that, for any n, our belief model for $Y_1, \ldots, Y_n$ is exchangeable:*

$$p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$$

for all partitions π of $\{1, \cdots, n\}$. Then our model can be written as

$$p(y_1, \ldots, y_n) = \int \left(\prod_{i=1}^{n} p(y_i \mid \theta) \right) p(\theta)\, d\theta.$$

for some parameter θ, some prior distribution on θ and some sampling model $p(y \mid \theta)$. The prior and sampling model depends on the form of the belief model $p(y_1, \cdots, y_n)$.

So, in general,

$$\left.\begin{array}{l} Y_1, \ldots, Y_n | \theta \text{ are i.i.d.} \\ \theta \sim p(\theta) \end{array}\right\} \Leftrightarrow Y_1, \ldots, Y_n \text{ are exchangeable for all } n.$$

Importantly, if we sample from a sufficiently large population, then we can model the sample variables as being approximately conditionally i.i.d.

3 One-parameter models

A one-parameter model is a class of sampling distributions that is indexed by a single unknown parameter.

Conjugate

Definition 3.1 *A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(y|\theta)$ if*

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}.$$

Some distributions

(1) Beta distribution, $\theta \sim \text{beta}(a, b)$

$$p(\theta) = \text{dbeta}(\theta, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad \text{for } 0 \leq \theta \leq 1.$$

$$\text{E}[\theta] = a/(a+b); \quad \text{Var}[\theta] = ab/[(a+b+1)(a+b)^2] = \text{E}[\theta] \times \text{E}[1-\theta]/(a+b+1).$$

(2) Binomial distribution, $Y \in \{0, 1, \ldots, n\} \sim \text{binomial}(n, \theta)$ distribution if:

$$\Pr(Y = y|\theta) = \text{dbinom}(y, n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y \in \{0, 1, \ldots, n\}.$$

$$\text{E}[Y|\theta] = n\theta; \quad \text{Var}[Y|\theta] = n\theta(1-\theta).$$

(3) Poisson distribution, a random variable Y has a Poisson distribution with mean θ

$$\Pr(Y = y|\theta) = \text{dpois}(y, \theta) = \frac{\theta^y e^{-\theta}}{y!} \quad \text{for} \quad y \in \{0, 1, 2, \ldots\}.$$

$$\text{E}[Y|\theta] = \theta; \quad \text{Var}[Y|\theta] = \theta.$$

"mean-variance" relationship

(4) Gamma distribution, $\theta \sim \text{gamma}(a, b)$

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

$$\text{E}[\theta] = a/b; \quad \text{Var}[\theta] = a/b^2$$

3.1 The Binomial model

Happiness data: $n = 129$, N: the total size of the female senior citizen population

$$Y_i = \begin{cases} 1 & \text{if Happy} \\ 0 & \text{if Unhappy} \end{cases}, \qquad \theta = \sum_{i=1}^{N} \frac{Y_i}{N}$$

(Likelihood function) The probability for any potential outcome $y_1, \ldots, y_{129}$, conditional on θ:

$$p(y_1, \ldots, y_{129}|\theta) = \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i}.$$

(Prior distribution) Uniform prior distribution

$$p(\theta) = 1 \text{ for all } \theta \in [0, 1].$$

(Posterior distribution)

$$\begin{aligned} p(\theta|y_1, \ldots, y_{129}) &= \frac{p(y_1, \ldots, y_{129}|\theta)p(\theta)}{p(y_1, \ldots, y_{129})} = p(y_1, \ldots, y_{129}|\theta) \times \frac{1}{p(y_1, \ldots, y_{129})} \\ &\propto p(y_1, \ldots, y_{129}|\theta). \end{aligned}$$

$$p(\theta|y_1, \ldots, y_{129}) = \theta^{118}(1-\theta)^{11} \times p(\theta)/p(y_1, \ldots, y_{129}) = \theta^{118}(1-\theta)^{11} \times 1/p(y_1, \ldots, y_{129}).$$

$1/p(y_1, \ldots, y_{129})$: the *scale* or *normalizing constant*

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

$$\begin{aligned}
1 &= \int_0^1 p(\theta|y_1,\ldots,y_{129})d\theta \\
1 &= \int_0^1 \theta^{118}(1-\theta)^{11}/p(y_1,\ldots,y_{129})d\theta \\
1 &= \frac{1}{p(y_1,\ldots,y_{129})}\int_0^1 \theta^{118}(1-\theta)^{11}d\theta \\
1 &= \frac{1}{p(y_1,\ldots,y_{129})}\frac{\Gamma(119)\Gamma(12)}{\Gamma(131)} \qquad \rightarrow p(y_1,\ldots,y_{129}) = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}
\end{aligned}$$

$$p(\theta|y_1,\ldots,y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{118}(1-\theta)^{11} = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{119-1}(1-\theta)^{12-1}.$$

3.1.1 Inference for exchangeable binary data

(1)Sufficient statistic

If $Y_1, \ldots, Y_n|\theta$ are i.i.d. binary(θ):

$$p(\theta|y_1, \ldots, y_n) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \times p(\theta)/p(y_1, \ldots, y_n).$$

Compare the relative probability of any two θ-values:

$$\begin{aligned}\frac{p(\theta_a|y_1, \ldots, y_n)}{p(\theta_b|y_1, \ldots, y_n)} &= \frac{\theta_a^{\sum y_i}(1-\theta_a)^{n-\sum y_i} \times p(\theta_a)/p(y_1, \ldots, y_n)}{\theta_b^{\sum y_i}(1-\theta_b)^{n-\sum y_i} \times p(\theta_b)/p(y_1, \ldots, y_n)} \\ &= \left(\frac{\theta_a}{\theta_b}\right)^{\sum y_i} \left(\frac{1-\theta_a}{1-\theta_b}\right)^{n-\sum y_i} \frac{p(\theta_a)}{p(\theta_b)}.\end{aligned}$$

- The probability density at θ_a relative to that at θ_b depends on $y_1, \ldots, y_n$ only through $\sum y_i$.

$\Rightarrow$ contains all the information about θ available from the data

$\Rightarrow$ $\sum y_i$, a **sufficient statistic** for θ and $p(y_1, \ldots, y_n|\theta)$.

(2)Posterior inference under a uniform prior distribution

- prior: $p(\theta) = 1$ or beta$(1, 1)$
- likelihood: $binomial(n, \theta)$
- posterior: $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{p(y)} = c(y)\theta^y(1-\theta)^{n-y}p(\theta) = c(y)\theta^y(1-\theta)^{n-y}$

$$1 = \int_0^1 c(y)\theta^y(1-\theta)^{n-y}d\theta = c(y)\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

$$\begin{aligned} p(\theta|y) &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y} = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1} \\ &= \text{beta}(y+1, n-y+1). \end{aligned}$$

(3)Posterior distributions under beta prior distributions

- prior: $\theta \sim \text{beta}(a, b)$
- likelihood: $Y|\theta \sim \text{binomial}(n, \theta)$
- posterior:

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{1}{p(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y}\theta^y(1-\theta)^{n-y} \\ &= c(n, y, a, b) \times \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\ &= \text{dbeta}(\theta, a+y, b+n-y) \sim \text{beta}(a+y, b+n-y). \end{aligned}$$

(4)Combining information

$$\mathrm{E}[\theta|y] = \frac{a+y}{a+b+n}, \mathrm{mode}[\theta|y] = \frac{a+y-1}{a+b+n-2}, \mathrm{Var}[\theta|y] = \frac{\mathrm{E}[\theta|y]\,\mathrm{E}[1-\theta|y]}{a+b+n+1}.$$

$$\begin{aligned}\mathrm{E}[\theta|y] &= \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\ &= \frac{a+b}{a+b+n} \times \text{prior expectation} + \frac{n}{a+b+n} \times \text{data average}.\end{aligned}$$

When $n \to \infty$ or $n >> a+b$,:

$$\frac{a+b}{a+b+n} \approx 0, \mathrm{E}[\theta|y] \approx \frac{y}{n}, \mathrm{Var}[\theta|y] \approx \frac{1}{n}\frac{y}{n}\left(1-\frac{y}{n}\right).$$

(5)Prediction

- **Posterior predictive** distribution, or sometimes **predictive distribution** is similar, but the expectation is taken with respect to the posterior

$$p(\tilde{x} \mid x) = \int_{\theta} f(\tilde{x} \mid \theta)\pi(\theta \mid x)d\theta \tag{2.12}$$

Notice that I distinguish $\tilde{x}$ and x on purpose to show that x is the "training data" while $\tilde{x}$ is the "testing data".

$\tilde{Y} \in \{0, 1\}$: An additional outcome from the same population that has yet to be observed.

The *predictive distribution* of $\tilde{Y}$ is

$$\begin{aligned}\Pr(\tilde{Y} = 1|y_1, \ldots, y_n) &= \int \Pr(\tilde{Y} = 1, \theta|y_1, \ldots, y_n)\, d\theta = \int \Pr(\tilde{Y} = 1|\theta, y_1, \ldots, y_n)p(\theta|y_1, \ldots, y_n)d\theta \\ &= \int \theta p(\theta|y_1, \ldots, y_n)\, d\theta = \mathrm{E}[\theta|y_1, \ldots, y_n] = \frac{a + \sum_{i=1}^{n} y_i}{a + b + n} \\ \Pr(\tilde{Y} = 0|y_1, \ldots, y_n) &= 1 - \mathrm{E}[\theta|y_1, \ldots, y_n] = \frac{b + \sum_{i=1}^{n}(1 - y_i)}{a + b + n}.\end{aligned}$$

3.1.2 Confidence regions

1. Bayesian coverage ($\sim$ credible intervals)

   Definition 3.2 *An interval $[l(y), u(y)]$, based on the observed data $Y = y$, has 95% Bayesian coverage for θ if*

   $$\Pr(l(y) < \theta < u(y)|Y = y) = 0.95$$

2. Frequentist coverage ($\sim$ confidence intervals)

   Definition 3.3 *A random interval $[l(Y), u(Y)]$ has 95% frequentist coverage for θ if, before the data are gathered,*

   $$\Pr(l(Y) < \theta < u(Y)|\theta) = 0.95$$

   In a sense, the frequentist and Bayesian notions of coverage describe pre- and post-experimental coverage, respectively.
   Can a confidence interval have the same Bayesian and frequentist coverage probability?
   An interval that has 95% Bayesian coverage additionally has the property that

   $$\Pr(l(Y) < \theta < u(Y)|\theta) = .95 + \epsilon_n$$

   where $|\epsilon_n| < a/n$ for some constant a.
   This means that a confidence interval procedure that gives 95% Bayesian coverage will have approximately 95% frequentist coverage as well, at least asymptotically

3. Quantile-based interval

   Goal: $100 \times (1-\alpha)\%$ quantile-based confidence interval (95%, $\alpha = 0.05 = 5\%$)

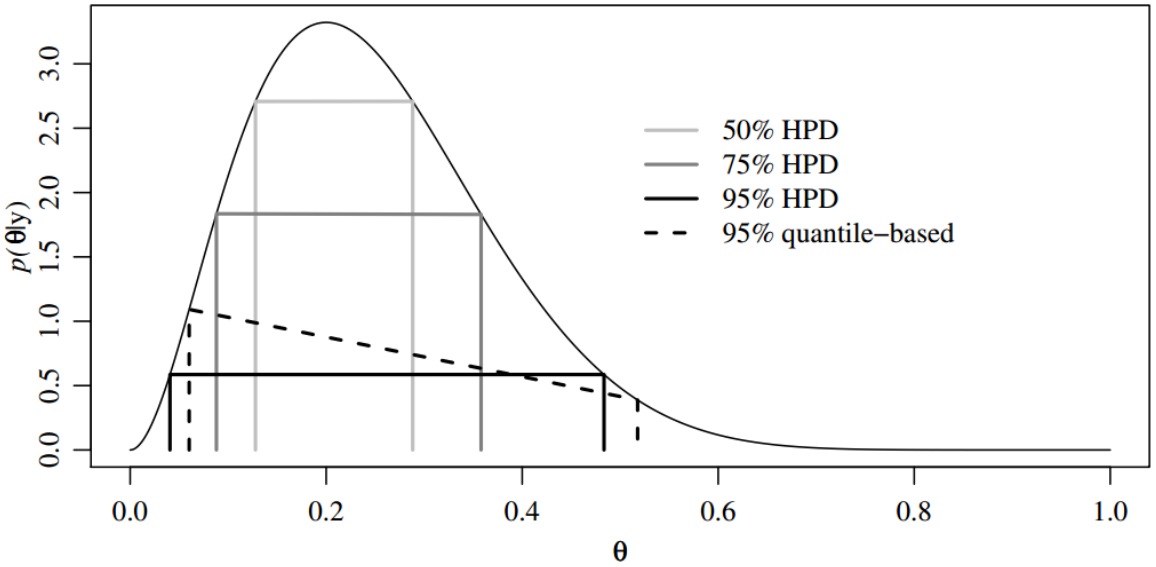
   How to do: Find numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$, the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of θ, such that

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

4. Highest posterior density (HPD) region

   Definition 3.4 *A $100 \times (1-\alpha)\%$ HPD region consists of a subset of the parameter space, $s(y) \subset \Theta$ such that $\Pr(\theta \in s(y)|Y = y) = 1-\alpha$; 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 an HPD region have a higher posterior density than points outside the region. However, an HPD region might not be an interval if the posterior density is multimodal (having multiple peaks).



Highest posterior density regions of varying probability content. The dashed line is the 95% quantile-based interval.

3.2 The Poisson model

3.2.1 Posterior inference

(1)Sufficient statistic

Model $Y_1, \ldots, Y_n$ as i.i.d. Poisson with mean θ, then the joint pdf of sample data:

$$\Pr(Y_1 = y_1, \ldots, Y_n = y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i}e^{-\theta} = c(y_1, \ldots, y_n)\theta^{\sum y_i}e^{-n\theta}.$$

Comparing two values of θ a *posteriori*:

$$\frac{p(\theta_a|y_1, \ldots, y_n)}{p(\theta_b|y_1, \ldots, y_n)} = \frac{c(y_1, \ldots, y_n)}{c(y_1, \ldots, y_n)} \frac{e^{-n\theta_a}}{e^{-n\theta_b}} \frac{\theta_a^{\sum y_i}}{\theta_b^{\sum y_i}} \frac{p(\theta_a)}{p(\theta_b)} = \frac{e^{-n\theta_a}}{e^{-n\theta_b}} \frac{\theta_a^{\sum y_i}}{\theta_b^{\sum y_i}} \frac{p(\theta_a)}{p(\theta_b)}.$$

As in the case of the i.i.d. binary model, $\sum_{i=1}^{n} Y_i$ contains all the information about θ that is available in the data, and again we say that $\sum_{i=1}^{n} Y_i$ is a sufficient statistic. Furthermore, $\{\sum_{i=1}^{n} Y_i\} \sim \text{Poisson}(n\theta)$

(2)Conjugate prior

Posterior distribution for θ: $p(\theta|y_1, \ldots, y_n) \propto p(\theta) \times p(y_1, \ldots, y_n|\theta) \propto p(\theta) \times \theta^{\sum y_i}e^{-n\theta}$

terms like $\theta^{c_1}e^{-c_2\theta} \Rightarrow$ the family of gamma distributions

(3)Posterior inference under a gamma prior distribution

Suppose $Y_1, \ldots, Y_n|\theta \sim$ i.i.d. Poisson(θ) and $p(\theta) = \text{dgamma}(\theta, a, b)$:

$$p(\theta|y_1, \ldots, y_n) = p(\theta) \times p(y_1, \ldots, y_n|\theta)/p(y_1, \ldots, y_n) = (\theta^{a-1}e^{-b\theta}) \times (\theta^{\sum y_i}e^{-n\theta}) \times c(y_1, \ldots, y_n, a, b)$$
$$= \left\{\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\right\} \times c(y_1, \ldots, y_n, a, b).$$

This is evidently a gamma distribution.

(4)Combining information

Posterior expectation of θ: a convex combination of the prior expectation and the sample average:

$$\text{E}[\theta|y_1, \ldots, y_n] = \frac{a + \sum y_i}{b + n} = \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n}$$

where b is the number of prior observations and a is the sum of counts from b prior observations.

$$n >> b \Rightarrow \text{E}[\theta|y_1, \ldots, y_n] \approx \bar{y}, \text{Var}[\theta|y_1, \ldots, y_n] \approx \bar{y}/n.$$

(5)Prediction

$$p(\theta) = \text{dgamma}(\theta, a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}, \Rightarrow \int_0^\infty \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}\, d\theta = 1 \Rightarrow \int_0^\infty \theta^{a-1}e^{-b\theta}\, d\theta = \frac{\Gamma(a)}{b^a}$$

Predictions about additional data can be obtained with the posterior predictive distribution:

$$\begin{aligned}
p(\tilde{y}|y_1,\ldots,y_n) &= \int_0^\infty p(\tilde{y}|\theta,y_1,\ldots,y_n)p(\theta|y_1,\ldots,y_n)d\theta = \int p(\tilde{y}|\theta)p(\theta|y_1,\ldots,y_n)\,d\theta \\
&= \int \text{dpois}(\tilde{y},\theta)\text{dgamma}(\theta,a+\sum y_i,b+n)d\theta \\
&= \int \left\{\frac{1}{\tilde{y}!}\theta^{\tilde{y}}e^{-\theta}\right\}\left\{\frac{(b+n)^{a+\sum y_i}}{\Gamma(a+\sum y_i)}\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\right\}d\theta \\
&= \frac{(b+n)^{a+\sum y_i}}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)}\int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1}e^{-(b+n+1)\theta}d\theta.
\end{aligned}$$

$$p(\tilde{y}|y_1,\ldots,y_n) = \left(\frac{b+n}{b+n+1}\right)^{a+\sum y_i}\left(\frac{1}{b+n+1}\right)^{\tilde{y}}\frac{\Gamma(a+\sum y_i+\tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)}$$

This is a negative binomial distribution with parameters $(a+\sum y_i, b+n)$, for which

$$\begin{aligned}
\text{E}[\tilde{Y}|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n} = \text{E}[\theta|y_1,\ldots,y_n]; \\
\text{Var}[\tilde{Y}|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n}\frac{b+n+1}{b+n} = \text{Var}[\theta|y_1,\ldots y_n]\times(b+n+1) = \text{E}[\theta|y_1,\ldots,y_n]\times\frac{b+n+1}{b+n}.
\end{aligned}$$

3.3 Exponential families and conjugate priors

A one-parameter exponential family model is any model whose densities can be expressed as

$$p(y|\phi) = h(y)c(\phi)e^{\phi t(y)},$$

where ϕ is the unknown parameter and $t(y)$ is the sufficient statistic.

- Prior distribution: $p(\phi|n_0, t_0) = \kappa(n_0, t_0)c(\phi)^{n_0}e^{n_0 t_0 \phi}$
- Likelihood: Information from $Y_1, \ldots, Y_n \sim$ i.i.d. $p(y|\theta)$

$\Downarrow$ Posterior distribution

$$\begin{aligned} p(\phi|y_1, \ldots, y_n) &\propto p(\phi)p(y_1, \ldots, y_n|\phi) \\ &\propto c(\phi)^{n_0+n} \exp\left\{\phi \times \left[n_0 t_0 + \sum_{i=1}^{n} t(y_i)\right]\right\} \\ &\propto p(\phi|n_0 + n, n_0 t_0 + n\bar{t}(\boldsymbol{y})) = p(\phi|n_0 + n, n_0 t_0 + \sum t(y_i)) \end{aligned}$$

- $n_0 \rightarrow$ "prior sample size" / a measure of how informative the prior is
  As a function of ϕ, $p(\phi|n_0, t_0)$ has the same shape as a likelihood $p(\tilde{y}_1, \ldots, \tilde{y}_{n_0}|\phi)$ based on n_0 "prior observations" $\tilde{y}_1, \ldots, \tilde{y}_{n_0}$ for which $\sum t(\tilde{y}_i)/n_0 = t_0$.. The prior distribution $p(\phi|n_0, t_0)$ contains the same amount of information that would be obtained from n_0 independent samples from the population.
- $t_0 \rightarrow$ "prior guess" of $t(Y)$ / the prior expected value of $t(Y)$

$$\mathrm{E}[t(Y)] = \mathrm{E}[\mathrm{E}[t(Y)|\phi]] = \mathrm{E}[-c'(\phi)/c(\phi)] = t_0$$

3.3.1 Example: Binomial model

$$p(y|\theta) = \theta^y(1-\theta)^{1-y} = \left(\frac{\theta}{1-\theta}\right)^y (1-\theta) = e^{\phi y}(1+e^{\phi})^{-1}.$$

The conjugate prior for ϕ is thus given by $p(\phi|n_0, t_0) \propto (1+e^{\phi})^{-n_0} e^{n_0 t_0 \phi}$, which can be translated into $p(\theta|n_0, t_0) \propto \theta^{n_0 t_0 - 1}(1-\theta)^{n_0(1-t_0)-1}$, a beta$(n_0 t_0, n_0(1-t_0))$ distribution.

The posterior would be $\{\theta|y_1, \ldots, y_n\} \sim$ beta$(t_0 + \sum y_i, (1-t_0) + \sum(1-y_i))$.

3.3.2 Example: Poisson model

The Poisson(θ) model can be shown to be an exponential family model with

- $t(y) = y$
- $\phi = \log\theta$
- $c(\phi) = \exp(e^{-\phi})$

The conjugate prior distribution for ϕ is thus $p(\phi|n_0, t_0) = \exp(n_0 e^{-\phi}) e^{n_0 t_0 y}$ where t_0 is the prior expectation of the population mean of Y. This translates into a prior density for θ of the form $p(\theta|n_0, t_0) \propto \theta^{n_0 t_0 - 1} e^{-n_0\theta}$, which is a gamma$(n_0 t_0, n_0)$ density. A weakly informative prior distribution can be obtained with t_0 set to the prior expectation of Y and $n_0 = 1$, giving a gamma $(t_0, 1)$ prior distribution. The posterior distribution under such a prior would be $\{\theta|y_1, \ldots, y_n\} \sim$ gamma $(t_0 + \sum y_i, 1+n)$

3.4 Discussion and supplement

3.4.1 Discussion

Most authors refer to intervals of high posterior probability as "credible intervals" as opposed to confidence intervals. Doing so fails to recognize that Bayesian intervals do have frequentist coverage probabilities, often being very close to the specified Bayesian coverage level.

3.4.2 Supplement

$$\text{if} \quad \left\{ \begin{array}{c} \theta \sim \text{beta}(1,1)\text{(uniform)} \\ Y \sim \text{binomial}(n,\theta) \end{array} \right\}, \quad \text{then} \quad \{\theta | Y = y\} \sim \text{beta}(1+y, 1+n-y)$$

$$\text{if} \quad \left\{ \begin{array}{c} \theta \sim \text{beta}(a,b) \\ Y \sim \text{binomial}(n,\theta) \end{array} \right\}, \quad \text{then} \quad \{\theta | Y = y\} \sim \text{beta}(a+y, b+n-y)$$

$$\text{if} \quad \left\{ \begin{array}{c} \theta \sim \text{gamma}(a,b) \\ Y_1, \ldots, Y_n | \theta \sim \text{Poisson}(\theta) \end{array} \right\}, \quad \text{then} \quad \{\theta | Y_1, \ldots, Y_n\} \sim \text{gamma}(a + \sum_{i=1}^{n} Y_i, b+n)$$

4 Monte Carlo approximation

- **Law of Large Number**
- **Central Limit Theorem**
  Under very general conditions, the sum (or mean) of a set of random variables is approximately normally distributed.

More than one parameter

$\rightarrow$ How to obtain exact values for these posterior quantities?

$\rightarrow$ Generate random sample values of the parameters from their posterior distributions by using MC

4.1 The Monte Carlo method

Monte Carlo approximation is based on random sampling and its implementation does not require a deep knowledge of calculus or numerical analysis.

4.1.1 Process

- θ: a parameter of interest; $y_1, \ldots, y_n$: the numerical values of **a sample** from a distribution $p(y_1, \ldots, y_n|\theta)$
- Sample some number S of independent, random θ-values from the posterior distribution $p(\theta|y_1, \ldots, y_n)$:

$$\theta^{(1)}, \ldots, \theta^{(S)} \sim \text{i.i.d } p(\theta|y_1, \ldots, y_n)$$

- the **empirical distribution** of the Monte Carlo samples $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \approx p(\theta|y_1, \ldots, y_n)$ especially with increasing S
- $$\left.\begin{array}{l} \text{sample } \theta^{(1)} \sim p(\theta|y_1, \ldots, y_n), \\ \text{sample } \theta^{(2)} \sim p(\theta|y_1, \ldots, y_n), \\ \qquad\qquad \vdots \\ \text{sample } \theta^{(S)} \sim p(\theta|y_1, \ldots, y_n), \end{array}\right\} \text{ independently .}$$

4.1.2 According to the law of large numbers

If $\theta^{(1)}, \ldots, \theta^{(S)}$ are i.i.d. samples from $p(\theta|y_1, \ldots, y_n)$, then

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to \mathrm{E}[g(\theta)|y_1, \ldots, y_n] = \int g(\theta)p(\theta|y_1, \ldots, y_n)d\theta \text{ as } S \to \infty$$

where $g(\theta)$ can be any function. Thus,

- $\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)} \longrightarrow \mathrm{E}[\theta|y_1, \ldots, y_n]$
- $\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2 \to \mathrm{Var}[\theta|y_1, \ldots, y_n]$
- $\#\frac{1}{S}(\theta^{(s)} \leq c) \to \Pr(\theta \leq c|y_1, \ldots, y_n)$
- the empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to p(\theta|y_1, \ldots, y_n)$
- the median of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{1/2}$
- the α-percentile of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_\alpha$

4.1.3 Monte Carlo standard errors

(assess the accuracy of approximations to posterior means):

$\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}$: the sample mean of the Monte Carlo samples,

- Monte Carlo estimate of expectation $\mathrm{E}[\theta|y_1, \ldots, y_n]$: $\bar{\theta}$ (approximately) **the Central Limit Theorem**
- Monte Carlo estimate of $\mathrm{Var}[\theta|y_1, \ldots, y_n]$: $\hat{\sigma}^2 = \frac{1}{S-1}\sum(\theta^{(s)} - \bar{\theta})^2$ **the Central Limit Theorem**
- Monte Carlo standard error: $\sqrt{\hat{\sigma}^2/S}$

- choose S so that $\sqrt{\hat{\sigma}^2/S}$ is less than the precision to which you want to report $\mathrm{E}[\theta|y_1,\ldots,y_n]$
- size $S = 100$, the estimate of $\mathrm{Var}[\theta|y_1,\ldots y_n] = 0.024$, thus $\sqrt{0.024/100} = 0.015$.
- The difference between $\mathrm{E}[\theta|y_1,\ldots,y_n]$ and its Monte Carlo estimate to be less than 0.01 with a high probability $\rightarrow 2\sqrt{0.024/S} < 0.01$, i.e. $S > 960$.

- An approximate 95% Monte Carlo confidence interval for the posterior mean of θ: $\hat{\theta} \pm 2\sqrt{\hat{\sigma}^2/S}$

4.2 Posterior inference for arbitrary functions

We may be interested in the posterior distribution of some computable function $g(\theta)$ of θ.

We have generated a sequence $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ from the posterior distribution of θ.

4.2.1 log odds function $\sim$ Binomial model

$$\gamma = \log \mathrm{odds}(\theta) = \log \frac{\theta}{1-\theta}.$$

- The average value of $\gamma = \log \frac{\theta}{1-\theta}$: converges to $\mathrm{E}[\log \frac{\theta}{1-\theta}|y_1,\ldots,y_n]$
  What about other aspects of posterior distribution of $\gamma = \log \frac{\theta}{1-\theta}$?

- using a Monte Carlo approach:

$$\left.\begin{array}{ll} \text{sample } \theta^{(1)} \sim p(\theta|y_1,\ldots,y_n), & \text{compute } \gamma^{(1)} = g(\theta^{(1)}) \\ \text{sample } \theta^{(2)} \sim p(\theta|y_1,\ldots,y_n), & \text{compute } \gamma^{(2)} = g(\theta^{(2)}) \\ \vdots & \\ \text{sample } \theta^{(S)} \sim p(\theta|y_1,\ldots,y_n), & \text{compute } \gamma^{(S)} = g(\theta^{(S)}) \end{array}\right\} \text{independently}$$

Sequence $\{\gamma^{(1)},\ldots,\gamma^{(S)}\}$ constitutes S independent samples from $p(\theta|y_1,\ldots,y_n)$, and so as $S \to \infty$

- $\bar{\gamma} = \frac{1}{S}\sum_{s=1}^{S}\gamma^{(s)} \to \mathrm{E}[\gamma|y_1,\ldots,y_n]$
- $\frac{1}{S-1}\sum_{s=1}^{S}(\gamma^{(s)}-\bar{\gamma})^2 \to \mathrm{Var}[\gamma|y_1,\ldots,y_n]$
- the empirical distribution of $\{\gamma^{(1)},\ldots,\gamma^{(S)}\} \to p(\gamma|y_1,\ldots,y_n)$

4.2.2 Functions of two parameters

$$\left.\begin{array}{lll} \text{sample } \theta_1^{(1)} \sim p(\theta_1|y_1,\ldots,y_n), & \text{sample } \theta_2^{(1)} \sim p(\theta_2|y_1,\ldots,y_n), & \text{compute } g(\theta_1^{(1)},\theta_2^{(1)}) \\ \text{sample } \theta_1^{(2)} \sim p(\theta_1|y_1,\ldots,y_n), & \text{sample } \theta_2^{(2)} \sim p(\theta_2|y_1,\ldots,y_n), & \text{compute } g(\theta_1^{(2)},\theta_2^{(2)}) \\ \vdots & \vdots & \vdots \\ \text{sample } \theta_1^{(S)} \sim p(\theta_1|y_1,\ldots,y_n), & \text{sample } \theta_2^{(S)} \sim p(\theta_2|y_1,\ldots,y_n), & \text{compute } g(\theta_1^{(S)},\theta_2^{(S)}) \end{array}\right\} \text{independently .}$$

Sequence $\{(\theta_1^{(1)},\theta_2^{(1)}),\ldots,(\theta_1^{(S)},\theta_2^{(S)})\}$ consists of S independent samples from the joint posterior distribution of θ_1 and θ_2, and can be used to make MC approximations to posterior quantities of interest.

- For example, $\Pr(\theta_1 > \theta_2 | \sum_{i=1}^{111} Y_{i,1} = 217, \sum_{i=1}^{44} Y_{i,2} = 66)$ is approximated by $\frac{1}{S}\sum_{s=1}^{S} 1(\theta_1^{(s)} > \theta_2^{(s)})$, where $1(x > y)$ is the indicator function which is 1 if $x > y$ and zero otherwise.

4.3 Sampling from predictive distributions

The (posterior) **predictive distribution** of a random variable $\tilde{Y}$ is a probability distribution for $\tilde{Y}$ such that

- known quantities have been conditioned on
- unknown quantities have been integrated out

4.3.1 Sampling from prior predictive distribution

If true θ is known, the **sampling model / likelihood function** about $\tilde{Y}$ is:

$$\Pr(\tilde{Y} = \tilde{y}|\theta) = p(\tilde{y}|\theta)$$

But θ is unknown $\rightarrow$ We cannot make predictions $\rightarrow$ integrate out θ, thus, the predictive model is:

$$\Pr(\tilde{Y} = \tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)d\theta$$

A predictive distribution that integrates over unknown parameters but is not conditional on observed data is called a prior predictive distribution.

4.3.2 Sampling from posterior predictive distribution

After observing a sample $Y_1, \ldots, Y_n$ from the population, the relevant predictive distribution:

$$\Pr(\tilde{Y} = \tilde{y} | Y_1 = y_1, \ldots, Y_n = y_n) = \int p(\tilde{y}|\theta, y_1, \ldots, y_n) p(\theta|y_1, \ldots, y_n) d\theta = \int p(\tilde{y}|\theta) p(\theta|y_1, \ldots, y_n)\, d\theta.$$

This is called a posterior predictive distribution, because it conditions on an observed dataset.

4.3.3 Sampling process

- Sample from $p(\theta|y_1, \ldots, y_n)$ ✓, sample from $p(y|\theta)$ ✓, but $p(\tilde{y}|y_1, \ldots, y_n)$ is too complicated to sample from directly.

⇓ Sample from the posterior predictive distribution indirectly using a Monte Carlo procedure, because:

$$p(\tilde{y}|y_1, \ldots, y_n) = \int p(\tilde{y}|\theta) p(\theta|y_1, \ldots, y_n) d\theta,$$

thus $p(\tilde{y}|y_1, \ldots, y_n)$ is the posterior expectation of $p(\tilde{y}|\theta)$.

Obtain the posterior predictive probability that $\tilde{Y}$ is equal to some specific value $\tilde{y}$:

Sample $\theta^{(1)}, \ldots, \theta^{(S)} \sim$ i.i.d. $p(\theta|y_1, \ldots, y_n)$, and approximate $p(\tilde{y}|y_1, \ldots, y_n)$ with $\sum_{s=1}^{S} p(\tilde{y}|\theta^{(s)})/S$, i.e.,

$$\begin{array}{ll} \text{sample } \theta^{(1)} \sim p(\theta|y_1, \ldots, y_n), & \text{sample } \tilde{y}^{(1)} \sim p(\tilde{y}|\theta^{(1)}) \\ \text{sample } \theta^{(2)} \sim p(\theta|y_1, \ldots, y_n), & \text{sample } \tilde{y}^{(2)} \sim p(\tilde{y}|\theta^{(2)}) \\ \vdots & \\ \text{sample } \theta^{(S)} \sim p(\theta|y_1, \ldots, y_n), & \text{sample } \tilde{y}^{(S)} \sim p(\tilde{y}|\theta^{(S)}). \end{array}$$

The sequence $\{(\theta, \tilde{y})^{(1)}, \ldots, (\theta, \tilde{y})^{(S)}\}$ constitutes S independent samples from the joint posterior distribution of $(\theta, \tilde{Y})$, and the sequence $\{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(S)}\}$ constitutes S independent samples from the marginal posterior distribution of $\tilde{Y}$, which is the posterior predictive distribution.

Once we have generated these Monte Carlo samples from the posterior predictive distribution, we can use them again to calculate other posterior quantities of interest.

4.3.4 Example

let $\tilde{Y}$ be the number of children of a person who is sampled from the population of women aged 40 with a college degree. θ, the mean birthrate of this population.

Sampling model: $\Pr(\tilde{Y} = \tilde{y}|\theta) = p(\tilde{y}|\theta) = \theta^{\tilde{y}} e^{-\theta}/\tilde{y}!$

Predictive model: $\Pr(\tilde{Y} = \tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)d\theta$

When $\theta \sim$gamma(a, b), this prior predictive distribution is the negative binomial(a, b) distribution.

posterior predictive distribution: $\Pr(\tilde{Y} = \tilde{y}|Y_1 = y_1, \ldots, Y_n = y_n) = \int p(\tilde{y}|\theta, y_1, \ldots, y_n)p(\theta|y_1, \ldots, y_n)d\theta =$

$\int p(\tilde{y}|\theta)p(\theta|y_1,\ldots,y_n)\,d\theta$, which is negative binomial$(a+\Sigma y_i, b+n)$.

Posterior predictive samples from the conjugate Poisson model can be generated as follows:

$$\begin{aligned}
&\text{sample } \theta^{(1)} \sim \text{gamma}(a+\textstyle\sum y_i, b+n), \quad \text{sample } \tilde{y}^{(1)} \sim \text{Poisson}(\theta^{(1)})\\
&\text{sample } \theta^{(2)} \sim \text{gamma}(a+\textstyle\sum y_i, b+n), \quad \text{sample } \tilde{y}^{(2)} \sim \text{Poisson}(\theta^{(2)})\\
&\qquad\qquad\qquad\vdots\\
&\text{sample } \theta^{(S)} \sim \text{gamma}(a+\textstyle\sum y_i, b+n), \quad \text{sample } \tilde{y}^{(S)} \sim \text{Poisson}(\theta^{(S)}).
\end{aligned}$$

4.4 Posterior predictive model checking

The sample of 40-year-old women without a college degree, $n = 111$

1. **Two conflict distributions:**
   - Empirical distribution: the number of women with exactly 2 children is 38, which is twice the number of women in the sample with 1 child
   - Posterior predictive distribution: the probability of sampling a woman with 2 children is slightly less probable than sampling a woman with 1
2. **Explanations:**
   - It is a result of sampling variability: The empirical distribution of sampled data does not generally match exactly the distribution of the population from which the data were sampled, and in fact may look quite different if the sample size is small.
     In such cases, having a predictive distribution that smoothes over the bumps of the empirical distribution may be desirable.
   - Sample data are correct. In contrast, the Poisson model is unable to represent this feature of the population because there is no Poisson distribution that has such a sharp peak at $y = 2$.
3. These explanations for the discrepancy between the empirical and predictive distributions can be assessed numerically with Monte Carlo simulation.
   (1) For every vector $\boldsymbol{y}$ of length $n = 111$, let $t(\boldsymbol{y})$ be the ratio of the number of 2's in $\boldsymbol{y}$ to the number of 1's, so for our observed data yobs, $t(\boldsymbol{y}_{obs}) = 2$.
   (2) Suppose we were to sample a different set of 111 women, obtaining a data vector $\tilde{\boldsymbol{Y}}$ of length 111 recording their number of children. We obtain Monte Carlo samples by:
   For each $s \in \{1, \ldots, S\}$,

- sample $\theta^{(s)} \sim p(\theta|\boldsymbol{Y} = \boldsymbol{y}_{\text{obs}})$
- sample $\tilde{\boldsymbol{Y}}^{(s)} = (\tilde{y}_1^{(s)}, \ldots, \tilde{y}_n^{(s)}) \sim$ i.i.d. $p(y|\theta^{(s)})$
- compute $t^{(s)} = t(\tilde{\boldsymbol{Y}}^{(s)})$

In this Monte Carlo sampling scheme,

- $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ are samples from the posterior distribution of θ
- $\{\tilde{\boldsymbol{Y}}^{(1)}, \ldots, \tilde{\boldsymbol{Y}}^{(S)}\}$ are posterior predictive datasets, each of size n
- $\{t^{(1)}, \ldots, t^{(S)}\}$ are samples from the posterior predictive distribution of $t(\tilde{\boldsymbol{Y}})$

4. **Result:** Out of 10,000 Monte Carlo datasets, only about a half of a percent had values of $t(\boldsymbol{y})$ that equaled or exceeded $t(\boldsymbol{y}_{obs})$

   Poisson model is flawed: We would hardly ever see a dataset that resembled our observed one in terms of $t(\boldsymbol{y})$.
5. **Model choice:**
   - In terms of data description, we should at least make sure that our model generates predictive datasets $\tilde{\boldsymbol{Y}}$ that resemble the observed dataset in terms of features that are of interest.
   - However, an incorrect model can still provide correct inference for some aspects of the true population, such as the sample mean and variance

4.5 Discussion

A common practice: using the posterior predictive distribution to assess model fit (Guttman (1967) and Rubin (1984)) $\rightarrow$ posterior predictive p-values (differ from p-values based on classical goodness-of-fit tests)

5 The normal model

- Discuss some of the properties of the normal distribution
- Show how to make posterior inference on the population mean and variance parameters
- Compare the sampling properties of the *standard Bayesian estimator of the population mean* to those of the *unbiased sample mean*
- Discuss the appropriateness of the normal model when the underlying data are not normally distributed.

The importance of the normal distribution stems primarily from the central limit theorem, which says that under very general conditions, the sum (or mean) of a set of random variables is approximately normally distributed. In practice, this means that the normal sampling model will be appropriate for data that result from **the additive effects of a large number of factors**.

5.1 The normal model

A random variable Y is said to be normally distributed with mean θ and variance $\sigma^2 > 0$ if the density of Y is given by:

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

- The dnorm, rnorm, pnorm, and qnorm commands in R take the standard deviation σ as their argument, not the variance σ^2.
- If $X \sim$ normal(μ,τ^2), $Y \sim$ normal(θ,σ^2) and X and Y are independent, then $aX + bY \sim$ normal$(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)$

5.2 Inference for the mean, conditional on the variance

Model: $\{Y_1, \ldots, Y_n | \theta, \sigma^2\} \sim$ i.i.d. normal (θ, σ^2)

Joint sampling density:

$$p(y_1, \ldots, y_n | \theta, \sigma^2) = \prod_{i=1}^{n} p(y_i | \theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \theta}{\sigma}\right)^2} = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\sum\left(\frac{y_i - \theta}{\sigma}\right)^2\right\}.$$

- Expanding the quadratic term in the exponent, $p(y_1, \ldots, y_n | \theta, \sigma^2)$ depends on $y_1, \ldots, y_n$ through

$$\sum_{i=1}^{n}\left(\frac{y_i - \theta}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum y_i^2 - 2\frac{\theta}{\sigma^2}\sum y_i + n\frac{\theta^2}{\sigma^2}$$

- Two-dimensional sufficient statistic: $\{\sum y_i^2, \sum y_i\}$; $\bar{y} = \sum y_i/n$, $s^2 = \sum(y_i - \bar{y})^2/(n-1)$, thus $\{\bar{y}, s^2\}$.

Problem: Inference for this two-parameter model $\xrightarrow{breakdown}$ Two one-parameter problems

Begin with making inference for θ when σ^2 is known, and use a conjugate prior distribution for θ.

For any (conditional) prior distribution $p(\theta|\sigma^2)$, the posterior distribution will satisfy:

$$\begin{aligned} p(\theta | y_1, \ldots, y_n, \sigma^2) &= p(\theta|\sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2)/p(y_1, \ldots, y_n|\sigma^2) \propto p(\theta|\sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2) \\ &\propto p(\theta|\sigma^2) \times e^{-\frac{1}{2\sigma^2}\sum(y_i - \theta)^2} \propto p(\theta|\sigma^2) \times e^{c_1(\theta - c_2)^2}. \end{aligned}$$

If $p(\theta|\sigma^2)$ conjugate $\rightarrow$ must include quadratic terms like $e^{c_1(\theta-c_2)^2} \rightarrow$ The simplest such class of probability

densities: the normal family of densities ⇓

If $p(\theta|\sigma^2)$ is normal and $y_1, \ldots, y_n$ are i.i.d. normal(θ, σ^2), then $p(\theta|y_1, \ldots, y_n, \sigma^2)$ is also a normal density.

Calculation Process If $\theta \sim$ normal(μ_0, τ_0^2), then

$$p(\theta|y_1, \ldots, y_n, \sigma^2) = p(\theta|\sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2)/p(y_1, \ldots, y_n|\sigma^2) \propto p(\theta|\sigma^2)p(y_1, \ldots, y_n|\theta, \sigma^2)$$
$$\propto \exp\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\} \exp\{-\frac{1}{2\sigma^2}\sum(y_i - \theta)^2\}.$$

Adding the terms in the exponents and ignoring the -1/2 for the moment:

$$\frac{1}{\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}(\sum y_i^2 - 2\theta\sum y_i + n\theta^2) = a\theta^2 - 2b\theta + c, \text{ where}$$

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \quad b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}, \quad \text{and } c = c(\mu_0, \tau_0^2, \sigma^2, y_1, \ldots, y_n).$$

$$p(\theta|\sigma^2, y_1, \ldots, y_n) \propto \exp\{-\frac{1}{2}(a\theta^2 - 2b\theta)\} = \exp\{-\frac{1}{2}a(\theta^2 - 2b\theta/a + b^2/a^2) + \frac{1}{2}b^2/a\}$$
$$\propto \exp\{-\frac{1}{2}a(\theta - b/a)^2\} = \exp\left\{-\frac{1}{2}\left(\frac{\theta - b/a}{1/\sqrt{a}}\right)^2\right\}.$$

the standard deviation: $1/\sqrt{a}$; the role of the mean: b/a. Thus, $p(\theta|\sigma^2, y_1, \ldots, y_n)$ is indeed a normal density, with mean μ_n and variance τ_n^2, where

$$\mu_n = \frac{b}{a} = (\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y})/(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}) \quad \text{and} \quad \tau_n^2 = \frac{1}{a} = 1/(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}).$$

5.2.1 Combining information

The (conditional) posterior parameters τ_n^2 and μ_n combine the prior parameters τ_0^2 and μ_0 from the data.

- Posterior variance and precision:

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2},$$

Inverse variance is often referred to as the precision. For the normal model, let: $\tilde{\sigma}^2 = 1/\sigma^2$= sampling precision, i.e. how close the y_i's are to θ; $\tilde{\tau}_0^2 = 1/\tau_0^2$= prior precision; $\tilde{\tau}_n^2 = 1/\tau_n^2$= posterior precision. Thus,

$$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2,$$

so posterior information = prior infor. + data infor., a larger sample size n increases this precision.

- Posterior mean

$$\mu_n = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\mu_0 + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\bar{y} = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}$$

the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$. (weight on the prior mean = $1/\tau_0^2$ = prior precision; weight on the sample mean = n/σ^2 = sampling precision)
If the prior mean were based on κ_0 prior observations from **the same (or similar)** population as $Y_1, \ldots, Y_n$, then we might want to set $\tau_0^2 = \sigma^2/\kappa_0$, the variance of the *mean* of the prior observations. Thus:

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} = \frac{\kappa_0}{\kappa_n}\mu_0 + \frac{n}{\kappa_n}\bar{y}. \quad (\text{Let} \kappa_n = \kappa_0 + n),$$

a weighted average of μ_0 and $\bar{y}$ given the number of prior "observations" κ_0 and the sample size n.

5.2.2 Prediction

Goal: Predict a new observation $\tilde{Y}$ from the population after having observed $(Y_1 = y_1, \ldots, Y_n = y_n)$. (find the predictive distribution)

A fact: saying that $\tilde{Y}$ is normal with mean θ is the same as saying $\tilde{Y}$ is equal to θ plus some mean-zero normally distributed noise, that is:

$$\{\tilde{Y}|\theta, \sigma^2\} \sim \text{normal}(\theta, \sigma^2) \Leftrightarrow \tilde{Y} = \theta + \tilde{\epsilon}, \quad \{\tilde{\epsilon}|\theta, \sigma^2\} \sim \text{normal}(0, \sigma^2).$$

First compute

$$\text{E}[\tilde{Y}|y_1, \ldots, y_n, \sigma^2] = \text{E}[\theta + \tilde{\epsilon}|y_1, \ldots, y_n, \sigma^2] = \text{E}[\theta|y_1, \ldots, y_n, \sigma^2] + \text{E}[\tilde{\epsilon}|y_1, \ldots, y_n, \sigma^2] = \mu_n + 0 = \mu_n$$

$$\text{Var}[\tilde{Y}|y_1, \ldots, y_n, \sigma^2] = \text{Var}[\theta + \tilde{\epsilon}|y_1, \ldots, y_n, \sigma^2] = \text{Var}[\theta|y_1, \ldots, y_n, \sigma^2] + \text{Var}[\tilde{\epsilon}|y_1, \ldots, y_n, \sigma^2] = \tau_n^2 + \sigma^2.$$

The predictive distribution is therefore $\tilde{Y}|\sigma^2, y_1, \ldots, y_n \sim \text{ normal}(\mu_n, \tau_n^2 + \sigma^2)$.

- uncertainty about the center of the population τ_n^2
  As $n \to \infty$, more and more certain about where θ is, and the posterior variance τ_n^2 of θ goes to zero.
- how variable the population σ^2
  But certainty about θ does not reduce the sampling variability σ^2

5.3 Joint inference for the mean and variance

For any joint prior distribution $p(\theta, \sigma^2)$ for θ and σ^2, posterior inference proceeds using Bayes' rule:

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

First, develop a simple conjugate class of prior distributions which makes posterior calculations easy.

- θ's prior distribution:
  - Recall1: a joint distribution for two quantities can be expressed as the product of a *conditional probability* and *a marginal probability*: $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$
  - Recall2: if σ^2 were known, then a conjugate prior distribution for θ was normal(μ_0, τ_0^2).
    - particular case: $\tau_0^2 = \sigma^2/\kappa_0$:

$$p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2) = \text{dnorm}(\theta, \mu_0, \tau_0 = \sigma/\sqrt{\kappa_0}) \times p(\sigma^2).$$

  μ_0 and κ_0 can be interpreted as the mean and sample size from a set of prior observations
- σ^2's prior distribution: the gamma family is a conjugate class of densities for $1/\sigma^2$ (precision). $\rightarrow \sigma^2$ has an *inverse-gamma* distribution: precision $= 1/\sigma^2 \sim$ gamma(a, b), $\sigma^2 \sim$ inverse-gamma(a, b). Rewrite as:

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

$\text{E}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$; $\quad \text{mode}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1}$, $\quad$ so mode$[\sigma^2] < \sigma_0^2 < \text{E}[\sigma^2]$; $\quad \text{Var}[\sigma^2]$ is decreasing in ν_0.
the prior parameters (σ_0^2, υ_0) as the sample variance and sample size of prior observations.

5.3.1 Posterior inference

Thus, our **prior distributions** and **likelihood functions** (sampling model) are as follows:

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

$$\theta|\sigma^2 \sim \text{normal}(\mu_0, \sigma^2/\kappa_0)$$

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

Just as the prior distribution for θ and σ^2 can be decomposed as $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$, the posterior distribution can be similarly decomposed:

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

- $p(\theta|\sigma^2, y_1, \ldots, y_n)$: can be obtained using the previous result: Plugging in σ^2/κ_0 for τ_0^2, thus

$$\{\theta|y_1, \ldots, y_n, \sigma^2\} \sim \text{normal}(\mu_n, \sigma^2/\kappa_n), \text{ where}$$

$$\kappa_n = \kappa_0 + n \text{ and } \mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}, \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2}{\kappa_0 + n}.$$

Therefore, if μ_0 is the mean of κ_0 prior observations, then $E[\theta|y_1, \ldots, y_n, \sigma^2]$ is the sample mean of the current and prior observations, and $Var[\theta|y_1, \ldots, y_n, \sigma^2]$ is σ^2 divided by the total number of observations, both prior and current.

- $p(\sigma^2|y_1, \ldots, y_n)$: can be obtained by performing an integration over the unknown value of θ:

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

The result is:

$$\begin{aligned} \{\sigma^2|y_1, \ldots, y_n\} &\sim \text{inverse} - \text{gamma}(\nu_n/2, \nu_n\sigma_n^2/2), \\ \{1/\sigma^2|y_1, \ldots, y_n\} &\sim \text{gamma}(\nu_n/2, \nu_n\sigma_n^2/2) \end{aligned}$$

where

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

ν_n is fairly intuitive, it acts as a sample size which is the "prior sample size" of the variance plus the sample size n. σ_n^2 is a bit harder to understand. There are three terms here.

1. $\nu_0\sigma_0^2$, can be thought of as a prior sum of squared observations from the sample mean (ν_0 prior samples with variance σ_0^2).
2. Similarly, $(n-1)s^2$, where $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$, is literally the sum of squared (actually observed) observations from the sample mean.
3. Lastly, the third term increases the posterior variance if the observed sample mean $(\bar{y})$ is *far away* from the expected prior mean μ_0, since this would suggest higher variance.
4. All three "sum of squares-ish" terms are combined, then divided by the total number of "observations" $\nu_n = n + \nu_0$, as commonly done to estimate variance from a sample.

5.3.2 Summary of posterior inference

1. In sum, for inference with the normal model, there are four prior parameters to specify:
   - σ_0^2, an initial estimate for the variance;
   - ν_0, a "prior sample size" from which the initial estimate of the *variance* is observed;
   - μ_0, an initial estimate for the population mean;
   - κ_0, a "prior sample size" from which the initial estimate of the *mean* is observed
2. Then we have
   - $1/\sigma^2 \sim \text{Gamma}(\nu_0/2, \sigma_0^2\nu_0/2)$
   - $\implies \mathbb{E}(\sigma^2) = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$ (use expectation of inverse gamma)
   - $\theta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$
   - $\implies \mathbb{E}(\theta) = \mu_0$
3. Now we have a new sample with n observations. Thus, the updated parameters are
   - $\nu_n = \nu_0 + n$
   - $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right]$
   - $\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$
   - $\kappa_n = \kappa_0 + n$
4. So that the posterior is finally
   - $1/\sigma^2 \mid y_1, \ldots, y_n \sim \text{Gamma}(\nu_n/2, \sigma_n^2\nu_n/2)$
   - Where $\mathbb{E}(\sigma^2 \mid y_1, \ldots, y_n) = \frac{\sigma_n^2\nu_n}{2(\nu_n/2-1)}$ (using the expectation of the inverse gamma)
   - $\theta \mid \sigma^2, y_1, \ldots, y_n \sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)$
   - Where $\mathbb{E}(\theta \mid y_1, \ldots, y_n, \sigma^2) = \mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$

   Note how the prior sample sizes for the variance and the mean are decoupled because they update differently.

However, it's common to set $\nu_0 = \kappa_0$.

5. The final posterior distribution can be obtained by:

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

5.3.3 Monte Carlo sampling

All we know (so far) is that (1) the conditional distribution of θ given the data and σ^2 is normal, and that (2) σ^2 given the data is inverse-gamma.

If we could generate marginal samples of θ, from $p(\theta|y_1, \ldots, y_n)$, then we could use the Monte Carlo method to approximate the above quantities of interest. $\rightarrow$ This is quite easy to do by generating samples of θ and σ^2 from their joint posterior distribution. $\rightarrow$ Using the following Monte Carlo procedure:

$$\begin{aligned} \sigma^{2(1)} &\sim \text{inverse gamma}(\nu_n/2, \sigma_n^2\nu_n/2), \quad \theta^{(1)} \sim \text{normal}(\mu_n, \sigma^{2(1)}/\kappa_n) \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \sigma^{2(S)} &\sim \text{inverse gamma}(\nu_n/2, \sigma_n^2\nu_n/2), \quad \theta^{(S)} \sim \text{normal}(\mu_n, \sigma^{2(S)}/\kappa_n). \end{aligned}$$

$\{\theta^{(n)}, \sigma^{2(n)}\}$ represent samples from the joint distribution $p(\theta, \sigma^2 \mid y_1, \ldots, y_n)$, and either set of values by themselves represents samples from the full marginal distribution. $\rightarrow \{\theta^{(1)}, \ldots, \theta^{(n)}\}$ can be seen as independent samples from the marginal posterior distribution of $p(\theta|y_1, \ldots, y_n)$

This is intuitive for $\sigma^{2(n)}$ but less so for $\theta^{(n)}$. The key is to notice that, although $\theta^{(n)}$ is sampled conditioned on $\sigma^{2(n)}$, multiple $\theta^{(n)}$ samples are conditioned on multiple *different* $\sigma^{2(n)}$s. Taken together, they constitute marginal samples of θ, and the $\theta^{(n)}$ do indeed represent samples from the marginal distribution.

5.3.4 Improper priors

What if we want to use *no* prior information? See what happens to our posterior distribution $\kappa_0, \nu_0 \to 0$.

$$\sigma_n^2 = \frac{1}{\nu_0 + n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2] \to \frac{n-1}{n}s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

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

.

Then, the "posterior" distribution (plugging in $\kappa_0 = \nu_0 = 0$ and the posterior parameters σ_n^2, μ_n and simplifying) would be

$$\{1/\sigma^2 | y_1, \ldots, y_n\} \sim \text{Gamma}(\frac{n}{2}, \frac{1}{n}\frac{n}{2}\sum(y_i - y)^2)$$
$$\{\theta | \sigma^2, y_1, \ldots, y_n\} \sim \mathcal{N}(\bar{y}, \frac{\sigma^2}{n})$$

With "significant algebra", you can show that inference this way results in

$$\frac{\theta - \bar{y}}{s/\sqrt{n}} | y_1, \ldots, y_n \sim t_{n-1}.$$

i.e. a t distribution with $n-1$ degrees of freedom. This is similar to the sampling distribution of t statistic:

$$\frac{\bar{Y} - \theta}{s/\sqrt{n}} | \theta \sim t_{n-1}.$$

But like the Bayesian vs Frequentist confidence intervals, they are philosophically different.

(1) describes uncertainty about the true mean conditional on the data (after you sample your data, your uncertainty is still represented with a t_{n-1} distribution)
(2) describes uncertainty about the observed sample mean given the true population mean (before you sample the data, your uncertainty about the scaled deviation of the sample mean $\tilde{Y}$ from the population mean θ is represented with a t_{n-1} distribution)

The difference is that before you sample your data, both $\tilde{Y}$ and θ are unknown. After sampling, $\tilde{Y} = \bar{y}$ is known, which provides us with information about θ. The improper prior on (θ, σ^2) lead to the above t_{n-1} posterior distribution for θ, and so inference based on this posterior distribution is not formally Bayesian.

5.4 Bias, variance and mean squared error

Now we are diving into the properties of estimators for posterior parameters.

A *point estimator* of an unknown parameter θ is a function that converts your data into a single element of the parameter space Θ. Good point estimators should hopefully approximate (and *reliably* approximate) the true value of θ; we can formalize these properties as the bias and mean squared error of estimators.

In Bayesian analysis, point estimators are usually functions of the posterior distribution of the parameter, such as the expectation. Suppose the true value of θ is θ_0.The *Bias* of an estimator $\hat{\theta}$ is

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta_0.$$

If $\text{Bias}(\hat{\theta}) = 0$ we say that $\hat{\theta}$ is an *unbiased* estimator; otherwise we say it is biased.

The point estimator for the posterior of our normal sampling model and a normal prior is (call it $\hat{\theta}_b$)

$$\hat{\theta}_b(y_1, \ldots, y_n) = \mathrm{E}[\theta | y_1, \ldots, y_n] = \frac{n}{\kappa_0 + n}\bar{y} + \frac{\kappa_0}{\kappa_0 + n}\mu_0 = w\bar{y} + (1 - w)\mu_0.$$

Consider the Bayesian $\hat{\theta}_b$ above V.S. the standard maximum likelihood estimator, $\hat{\theta}_e(y_1, \ldots, y_n) = \bar{y}$.

- $\mathrm{Bias}(\hat{\theta}_e) = \mathrm{E}[\hat{\theta}_e | \theta = \theta_0] - \theta_0 = \mathbb{E}(\hat{\theta}_e) - \theta_0 = 0$, so $\hat{\theta}_b$ is unbiased;
- $\mathrm{Bias}(\hat{\theta}_b) = \mathrm{E}[\hat{\theta}_b | \theta = \theta_0] - \theta_0 = \mathbb{E}(\hat{\theta}_b) - \theta_0 = w\theta_0 + (1 - w)\mu_0 - \theta$.
  Notice that the first two terms add up to θ only if $\mu_0 = \theta$. For all $\mu \neq \theta_0$, $\hat{\theta}_b$ is biased!

A biased estimator seems undesirable, but can actually be useful in this setting. Imagine "biasing" the estimator *towards the true mean* to obtain a more accurate estimate. Thus it is useful to recall using the Mean Squared Error as another measure of estimator performance, which measures how close an estimator $\hat{\theta}$ will be to the true population parameter θ, on average:

The *Mean Squared Error* (MSE) of an estimator $\hat{\theta}$ is

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta}) \text{or} = \mathrm{Var}[\hat{\theta}|\theta_0] + \mathrm{Bias}^2[\hat{\theta}|\theta_0].$$

This formulation is obtained by: Letting $m = E[\hat{\theta}|\theta_0]$, the MSE is

$$\begin{aligned}\mathrm{MSE}[\hat{\theta}|\theta_0] &= \mathrm{E}[(\hat{\theta} - \theta_0)^2|\theta_0] = \mathrm{E}[(\hat{\theta} - m + m - \theta_0)^2|\theta_0] \\ &= \mathrm{E}[(\hat{\theta} - m)^2|\theta_0] + 2\,\mathrm{E}[(\hat{\theta} - m)(m - \theta_0)|\theta_0] + \mathrm{E}[(m - \theta_0)^2|\theta_0].\end{aligned}$$

Since $m = E[\hat{\theta}|\theta_0]$ it follows that $E[\hat{\theta} - m|\theta_0] = 0$ and so the second term is zero, the formulation ✓.

$$\text{Var}[\hat{\theta}_e|\theta = \theta_0, \sigma^2] = \textbf{Var}(\bar{y}) = \frac{\sigma^2}{n},$$

$$\text{Var}[\hat{\theta}_b|\theta = \theta_0, \sigma^2] = \textbf{Var}(w\theta_0 + (1-w)\mu_0) = \textbf{Var}(w\theta_0) = w^2 \times \frac{\sigma^2}{n} < \frac{\sigma^2}{n}$$

Thus,

$$\text{MSE}[\hat{\theta}_e|\theta_0] = E[(\hat{\theta}_e - \theta_0)^2|\theta_0] = \text{Var}[\hat{\theta}_e|\theta = \theta_0, \sigma^2] + 0 = \frac{\sigma^2}{n}$$

$$\begin{aligned}
\text{MSE}[\hat{\theta}_b|\theta_0] &= E[(\hat{\theta}_b - \theta_0)^2|\theta_0] == E[\{w(\bar{y} - \theta_0) + (1-w)(\mu_0 - \theta_0)\}^2|\theta_0] \\
&= \textbf{Var}(\hat{\theta}_b) + \textbf{Bias}^2(\hat{\theta}_b) \\
&= w^2\frac{\sigma^2}{n} + [w\theta + (1-w)\mu_0 - \theta]^2 = w^2\frac{\sigma^2}{n} + [(1-w)\mu_0 - (1-w)\theta]^2 \\
&= w^2\frac{\sigma^2}{n} + (1-w)^2(\mu_0 - \theta)^2
\end{aligned}$$

Notice that

$$\text{MSE}(\hat{\theta}_b) < \text{MSE}(\hat{\theta}_e) \quad \text{if}$$

$$\begin{aligned}
&\implies w^2\frac{\sigma^2}{n} + (1-w)^2(\mu_0-\theta)^2 < \frac{\sigma^2}{n} \quad \implies (1-w)^2(\mu_0-\theta)^2 < (1-w^2)\frac{\sigma^2}{n} \\
&\implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{1-w^2}{(1-w)^2} \quad \implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{(1-w)(1+w)}{(1-w)^2} \\
&\implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{1+w}{1-w} \quad \implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{1+\frac{n}{\kappa_0+n}}{1-\frac{n}{\kappa_0+n}} \\
&\implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{\frac{\kappa_0+2n}{\kappa_0+n}}{\frac{\kappa_0}{\kappa_0+n}} \quad \implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{\kappa_0+2n}{\kappa_0+n}\frac{\kappa_0+n}{\kappa_0} \\
&\implies (\mu_0-\theta)^2 < \frac{\sigma^2}{n}\frac{\kappa_0+2n}{\kappa_0} \quad \implies (\mu_0-\theta)^2 < \sigma^2\frac{\kappa_0+2n}{n\kappa_0} \\
&\implies (\mu_0-\theta)^2 < \sigma^2\left(\frac{\kappa_0}{n\kappa_0} + \frac{2n}{n\kappa_0}\right) \quad \implies (\mu_0-\theta)^2 < \sigma^2\left(\frac{1}{n} + \frac{2}{\kappa_0}\right)
\end{aligned}$$

$$(\mu_0-\theta_0)^2 < \frac{\sigma^2}{n}\frac{1+w}{1-w} = \sigma^2\left(\frac{1}{n} + \frac{2}{\kappa_0}\right).$$

So the Bayesian estimator has lower mean squared error than the ML estimate as long as values of μ_0 and κ_0 are picked such that this inequality holds-intuitively, if your "guess" about the prior is not far from the truth.

5.5 Prior specification based on expectations

A p-dimensional exponential family model is a model whose densities can be written as $p(y|\boldsymbol{\phi}) = h(y)c(\boldsymbol{\phi})\exp\{\boldsymbol{\phi}^T\boldsymbol{t}(y)\}$, where $\boldsymbol{\phi}$ is the parameter to be estimated and $\boldsymbol{t}(y) = \{t_1(y), \ldots, t_p(y)\}$ is the sufficient statistic.

The normal model is a two-dimensional exponential family model. The normal density:

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

The exponential family parameters:

- $\boldsymbol{t}(y) = (y, y^2)$,
- $\phi = (\theta/\sigma^2, -(2\sigma^2)^{-1})$
- $c(\phi) = |\phi_2|^{1/2}\exp\{\phi_1^2/(2\phi_2)\}$.
- $h(y) = 1/\sqrt{\pi}$

Reconstruct the normal density:

$$\begin{aligned} p(y \mid \boldsymbol{\phi}) &= \frac{1}{\sqrt{\pi}}|\phi_2|^{1/2}\exp\left(\frac{\phi_1^2}{2\phi_2}\right)\exp\left(\begin{pmatrix} y & y^2 \end{pmatrix}\begin{pmatrix} \theta/\sigma^2 \\ -(2\sigma^2)^{-1} \end{pmatrix}\right) \\ &= \frac{1}{\sqrt{\pi}}(2\sigma^2)^{-1/2}\exp\left(\frac{(\theta/\sigma^2)^2}{-2(2\sigma^2)^{-1}}\right)\exp\left(\begin{pmatrix} y & y^2 \end{pmatrix}\begin{pmatrix} \theta/\sigma^2 \\ -(2\sigma^2)^{-1} \end{pmatrix}\right) \end{aligned}$$

I am not going to do the exact algebra here, but notice that once you combine the exponential terms (and the matrix multiplication in the second exp), there are three separate terms added together. With a common factor of $1/-2\sigma^2$, those three terms are the y^2, $2\theta y$, and θ^2 of the expanded normal density above.

With exponential family models, we can now "read off" conjugate priors; for the p-dimensional case, the prior is $p(\boldsymbol{\phi} \mid n_0, \mathbf{t}_0) \propto c(\boldsymbol{\phi})^{n_0} \exp(n_0 \mathbf{t}_0^T \boldsymbol{\phi})$. Using the change of variables formula (which seems very complicated), you can reparamaterize the corresponding prior in terms of θ and σ^2, which gives a prior that is the product of two priors we had determined previously to be conjugate: the normal and inverse-gamma densities.

There are some more details on the significance of specifying n_0 and $\mathbf{t}_0$ that I am skipping since it essentially mirrors the prior specification advice in the previous sections.

5.6 The normal model for non-normal data

Because of the central limit theorem (while the sampling distribution of a single data point is not normal, the sampling distribution of the sample mean is close to normal) etc., we often use the normal model for non-normal data. This is especially applicable when

1) We are measuring summary statistics of a population, such as the mean
2) When we are measuring variables that might be the additive result of many underlying factors, which results in an approximately normal variable

The central limit theorem tells us that

$$p(\bar{y}|\theta, \sigma^2) \approx \text{dnorm}(\bar{y}, \theta, \sqrt{\sigma^2/n}),$$

with the approximation becoming increasingly good as n gets larger.

If the population variance σ^2 were known, then an approximate posterior distribution of the population mean θ, conditional on the sample mean $\bar{y}$, could be obtained as

$$p(\theta|\bar{y}, \sigma^2) \propto p(\theta) \times p(\bar{y}|\theta, \sigma^2) \approx p(\theta) \times \text{dnorm}(\bar{y}, \theta, \sqrt{\sigma^2/n}).$$

Of course, σ^2 is generally not known, but it is estimated by s^2. The approximate posterior distribution of (θ, σ^2) conditional on the estimates $(\bar{y}, s^2)$ is given by

$$\begin{aligned} p(\theta, \sigma^2|\bar{y}, s^2) &\propto p(\theta, \sigma^2) \times p(\bar{y}, s^2|\theta, \sigma^2) \\ &= p(\theta, \sigma^2) \times p(\bar{y}|\theta, \sigma^2) \times p(s^2|\bar{y}, \theta, \sigma^2) \\ &\approx p(\theta, \sigma^2) \times \text{dnorm}(\bar{y}, \theta, \sqrt{\sigma^2/n}) \times p(s^2|\bar{y}, \theta, \sigma^2). \end{aligned} \tag{5.1}$$

Again, for large n, the approximation of $p(\bar{y}|\theta, \sigma^2)$ by the normal density is generally a good one even if the population is not normally distributed. However, it is not clear what to put for $p(s^2|\bar{y}, \theta, \sigma^2)$.

If we knew that the data were actually sampled from a normal distribution, then results from statistical theory would say that

$$p(s^2|\bar{y}, \theta, \sigma^2) = \text{dgamma}(s^2, \frac{n-1}{2}, \frac{n-1}{2\sigma^2}).$$

However, if the data are not normally distributed, then s^2 is not necessarily gamma-distributed or independent of $\bar{y}$. The use of the posterior distribution in Eq.(5.1) for non-normal data could give misleading results about the joint distribution of $\{\theta, \sigma^2\}$.

However, the marginal posterior distribution of θ based on Eq.(5.1) can be remarkably accurate, even for non-normal data. The reasoning is as follows: The central limit theorem says that for large n

$$\sqrt{n}\frac{\bar{Y}-\theta}{\sigma} \dot{\sim} \text{normal}(0,1),$$

where $\dot{\sim}$ means "approximately distributed as." Additionally, if n is sufficiently large, then $s^2 \approx \sigma^2$ and so

$$\sqrt{n}\frac{\bar{Y}-\theta}{s} \dot{\sim} \text{normal}(0,1).$$

This should seem familiar: Recall from introductory statistics that for normal data, $\sqrt{n}\frac{\bar{Y}-\theta}{s}$ has a t-distribution with $n-1$ degrees of freedom. For large n, s^2 is very close to σ^2 and the t_{n-1} distribution is very close to a normal$(0,1)$ distribution.

Even though the posterior distribution based on a normal model may provide good inference for the population mean, the normal model can provide misleading results for other sample quantities.

5.7 Discussion and further references

Justify the normal sampling model:

- Among all distributions with a given mean θ and variance σ^2, the normal(θ, σ^2) distribution is the most diffuse in terms of a measure known as entropy
- data analysis perspective
  - The sample mean will generally be approximately normally distributed due to the central limit theorem. Thus the normal model provides a reasonable sampling model for the sample mean, if not the sample data
  - The normal model is a simple exponential family model with sufficient statistics equivalent to the sample mean and variance. As a result, it will provide a consistent estimation of the population mean and variance even if the underlying population is not normal
  - Confidence intervals for the population mean based on the normal model will generally be asymptotically correct

6 Posterior approximation with the Gibbs sampler

It is easy to sample from the full conditional distribution of each parameter instead of the joint posterior distribution. Thus, posterior approximation can be made with the Gibbs sampler, an iterative algorithm that constructs a dependent sequence of parameter values whose distribution converges to the target joint posterior distribution.

6.1 A semiconjugate prior distribution

In Chapter 5, we performed two-parameter inference by decomposing the prior $p(\theta, \sigma^2) = p(\theta \mid \sigma^2)p(\sigma^2)$. So our prior distribution on θ depending on σ^2:

$$\theta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$$

Sometimes we may want to specify our uncertainty about θ as being independent of σ^2, so that $p(\theta, \sigma^2) = p(\theta)p(\sigma^2)$. One such joint distribution is the following "semiconjugate" prior distribution:

$$\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$$

$$1/\sigma^2 \sim \text{Gamma}(v_0/2, v_0\sigma_0^2/2)$$

If τ_0^2 is not proportional to σ^2, the marginal density of $1/\sigma^2$ is not a gamma distribution or any other standard distribution from which we can easily sample.

Why posterior densities are hard to calculate for nonconjugate priors using the Bayes rule?

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'}$$

The numerator here is often easy to calculate, but the denominator is often prohibitively hard to compute. When the numerator is not known to be proportional to a known probability distribution, we can't get the full joint density without the denominator.

However, consider that we may want to decouple the priors of the two parameters. This allows flexibility with the specification of the prior (initial estimate and confidence) of either parameter.

Consider the midge wing example: we picked a prior on θ that was centered around 1.9 (our prior expectation) but with most of its mass above 0, since wing lengths cannot be above 0. We can't freely do this from what we know in section 5 (i.e. setting $\tau_0^2 = \sigma^2/\kappa_0$). Alternatively, we can set τ_0^2 to be whatever we want, but then there is no longer a known form of the joint posterior

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

that can easily be sampled from. However, as it turns out, the full conditionals $p(\theta \mid \sigma^2, y_1, \ldots, y_n)$ and $p(\sigma^2 \mid \theta, y_1, \ldots, y_n)$ are easy to specify, as when evaluating the formulas, we can simply disregard the other fixed parameter as a constant, leading to known posterior distributions. A technique called Gibbs sampling allows us to take advantage of this by constructing a sampler that approximates the (unknown) joint distribution by sampling iteratively from the (known) full conditional distributions.

6.2 Discrete approximations

(Some basic setting / information):

The precision: $\tilde{\sigma}^2 = 1/\sigma^2$. The joint distribution (built out of standard prior and sampling distributions):

$$\begin{aligned} p(\theta, \tilde{\sigma}^2, y_1, \ldots, y_n) &= p(\theta, \tilde{\sigma}^2) \times p(y_1, \ldots, y_n | \theta, \tilde{\sigma}^2) \\ &= \text{dnorm}(\theta, \mu_0, \tau_0) \times \text{dgamma}(\tilde{\sigma}^2, \nu_0/2, \nu_0\sigma_0^2/2) \times \prod_{i=1}^{n} \text{dnorm}(y_i, \theta, 1/\sqrt{\tilde{\sigma}^2}). \end{aligned} \tag{6.1}$$

A discrete approximation to the posterior distribution: constructing a posterior distribution over a grid of parameter values, based on relative posterior probabilities. This is done by evaluating $p(\theta, \tilde{\sigma}^2, y_1, \ldots, y_n)$ on a two-dimensional grid of values of $\{\theta, \tilde{\sigma}^2\}$.

Letting $\{\theta_1, \ldots, \theta_G\}$ and $\{\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_H^2\}$ be sequences of evenly spaced parameter values, the discrete approximation to the posterior distribution assigns a posterior probability to each pair $\{\theta_k, \tilde{\sigma}_l^2\}$ on the grid, which is a real joint probability distribution for $\theta \in \{\theta_1, \ldots, \theta_G\}$ and $\tilde{\sigma} \in \{\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_H^2\}$

$$\begin{aligned} p_D(\theta_k, \tilde{\sigma}_l^2 | y_1, \ldots, y_n) &= \frac{p(\theta_k, \tilde{\sigma}_l^2 | y_1, \ldots, y_n)}{\sum_{g=1}^{G} \sum_{h=1}^{H} p(\theta_g, \tilde{\sigma}_h^2 | y_1, \ldots, y_n)} \\ &= \frac{p(\theta_k, \tilde{\sigma}_l^2, y_1, \ldots, y_n)/p(y_1, \ldots, y_n)}{\sum_{g=1}^{G} \sum_{h=1}^{H} p(\theta_g, \tilde{\sigma}_h^2, y_1, \ldots, y_n)/p(y_1, \ldots, y_n)} \\ &= \frac{p(\theta_k, \tilde{\sigma}_l^2, y_1, \ldots, y_n)}{\sum_{g=1}^{G} \sum_{h=1}^{H} p(\theta_g, \tilde{\sigma}_h^2, y_1, \ldots, y_n)}. \end{aligned}$$

in the sense that it sums to 1: $\sum_{g=1}^{G}\sum_{h=1}^{H} p(\theta_g, \tilde{\sigma}_h^2|y_1, \ldots, y_n) = 1$

According to Eq.6.1, we can obtain the $p_D(\theta_k, \tilde{\sigma}_l^2|y_1, \ldots, y_n)$. In general, to construct a similarly fine approximation for a p-dimensional posterior distribution we would need a p-dimensional grid containing 100^p posterior probabilities. This means that discrete approximations will only be feasible for densities having a small number of parameters.

Besides, according to Eq.6.1, the relative posterior probabilities of one set of parameter values $\{\theta_1, \tilde{\sigma}_1^2\}$ to another $\{\theta_2, \tilde{\sigma}_2^2\}$ can be directly computable as their ratio cancels out the integral:

$$\frac{p(\theta_1, \tilde{\sigma}_1^2|y_1, \ldots, y_n)}{p(\theta_2, \tilde{\sigma}_2^2|y_1, \ldots, y_n)} = \frac{p(\theta_1, \tilde{\sigma}_1^2, y_1, \ldots, y_n)/p(y_1, \ldots, y_n)}{p(\theta_2, \tilde{\sigma}_2^2, y_1, \ldots, y_n)/p(y_1, \ldots, y_n)} = \frac{p(\theta_1, \tilde{\sigma}_1^2, y_1, \ldots, y_n)}{p(\theta_2, \tilde{\sigma}_2^2, y_1, \ldots, y_n)}.$$

6.3 Sampling from the conditional distributions

To proceed with Gibbs sampling, need to calculate the full conditional distributions of the parameters.

6.3.1 The full conditional distribution

1. (In Chapter 5) For $\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$, then $\theta|\sigma^2, y_1, \ldots, y_n \sim \mathcal{N}(\mu_n, \tau_n^2)$: $\theta|\sigma^2, y_1, \ldots, y_n$

$$\{\sigma^2, \theta|y_1, \ldots, y_n\} \sim \text{normal}(\mu_n, \sigma^2/\kappa_n), \text{ where}$$

$$\kappa_n = \kappa_0 + n \text{ and } \mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}, \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2}{\kappa_0 + n}.$$

2. $\tilde{\sigma}^2|\theta, y_1, \ldots, y_n$

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

If θ and $\tilde{\sigma}^2$ are independent in the prior distribution, then $p(\theta|\tilde{\sigma}^2) = p(\theta)$:

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

So $\sigma^2 \sim$ inverse-gamma$(\nu_n/2, \nu_n\sigma_n^2(\theta)/2)$, and $\tilde{\sigma}^2 \sim$ gamma$(\nu_n/2, \nu_n\sigma_n^2(\theta)/2))$, with the parameters:

$$\nu_n = \nu_0 + n\ , \quad \sigma_n^2(\theta) = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + ns_n^2(\theta)\right]\ , s_n^2(\theta) = \sum(y_i-\theta)^2/n$$

$s_n^2(\theta)$ is the unbiased estimate of σ^2 if θ were known; denote $\sigma_n^2(\theta)$, $s_n^2(\theta)$ to indicate that σ_n^2 is dependent on θ which is assumed known.
Now, sample directly from $p(\sigma^2|\theta, y_1, \ldots, y_n)$ ✓; sample directly from $p(\theta|\sigma^2, y_1, \ldots, y_n)$ ✓; BUT do not have a way to sample directly from $p(\sigma^2, \theta|y_1, \ldots, y_n)$.

3. Use the full conditional distributions to sample from the joint posterior distribution
   Given $\sigma^{2(1)}$ (from the marginal posterior distribution $p(\sigma^2|\theta, y_1, \ldots, y_n)$)
   $\rightarrow$ Sample $\theta^{(1)} \sim p(\theta|\sigma^{2(1)}, y_1, \ldots, y_n)$ ($\theta^{(1)}$, a sample from the marginal distribution $p(\theta|\sigma^2, y_1, \ldots, y_n)$; $\{\theta^{(1)}, \sigma^{2(1)}\}$, a sample from the joint distribution of $\{\theta, \sigma^2\}$)
   $\rightarrow$ Sample $\sigma^{2(2)} \sim p(\sigma^2|\theta^{(1)}, y_1, \ldots, y_n)$ ($\sigma^{2(2)}$, a sample from the marginal distribution $p(\sigma^2|y_1, \ldots, y_n)$; $\{\theta^{(1)}, \sigma^{2(2)}\}$, a sample from the joint distribution of $\{\theta, \sigma^2\}$) $\rightarrow \ldots$
   Two conditional distributions could be used to generate samples from the joint distribution if only we had a $\sigma^{2(1)}$ from which to start
   The distributions $p(\theta|\sigma^2, y_1, \ldots, y_n)$ and $p(\sigma^2|\theta, y_1, \ldots, y_n)$ are called the *full conditional distributions* of θ and σ^2 respectively, as they are each a conditional distribution of a parameter given everything else.

6.4 Gibbs sampling

Given a current state of the parameters $\phi^{(s)} = \{\theta^{(s)}, \tilde{\sigma}^{2(s)}\}$, we generate a new state as follows:

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

This algorithm is called the Gibbs sampler and generates a dependent sequence of our parameters $\{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(S)}\}$.

6.5 General properties of the Gibbs sampler

Goal: $p(\boldsymbol{\phi}|y_1, \ldots, y_n)$

Begin with start values $\phi_i^{(0)}$ for all i. (only have to have start values for *all but one* parameter)

- Sample $\phi_1^{(1)} \sim p(\phi_1|\phi_2^{(0)}, \ldots, \phi_n^{(0)}, \ldots)$ (the full conditional distribution)
- Similarly, sample $\phi_2^{(1)} \sim p(\phi_2|\phi_1^{(1)}, \ldots, \phi_n^{(0)}, \ldots)$, $\phi_3^{(1)} \sim p(\phi_3|\phi_1^{(1)}, \phi_2^{(1)} \ldots, \phi_n^{(0)}, \ldots)$.
- $\boldsymbol{\phi}^{(1)} = \begin{pmatrix} \phi_1^{(1)} & \cdots & \phi_n^{(1)} \end{pmatrix}$ is your first Gibbs sample.

Thus, given starting point $\boldsymbol{\phi}^{(0))} = \{\phi_1^{(0))}, \ldots, \phi_p^{(0))}\}$, the Gibbs sampler generates $\boldsymbol{\phi}^{(s))}$ from $\boldsymbol{\phi}^{(s-1))}$ as follows:

$$
\begin{aligned}
&1.\ \text{sample } \phi_1^{(s)} \sim p(\phi_1|\phi_2^{(s-1)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)}) \\
&2.\ \text{sample } \phi_2^{(s)} \sim p(\phi_2|\phi_1^{(s)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)}) \\
&\vdots \\
&p.\ \text{sample } \phi_p^{(s)} \sim p(\phi_p|\phi_1^{(s)}, \phi_2^{(s)}, \ldots, \phi_{p-1}^{(s)}).
\end{aligned}
$$

This algorithm generates a *dependent* sequence of vectors:

$$
\begin{aligned}
\boldsymbol{\phi}^{(1)} &= \{\phi_1^{(1)}, \ldots, \phi_p^{(1)}\} \\
\boldsymbol{\phi}^{(2)} &= \{\phi_1^{(2)}, \ldots, \phi_p^{(2)}\} \\
\boldsymbol{\phi}^{(S)} &= \left\{\phi_1^{(S)}, \ldots, \phi_p^{(S)}\right\}.
\end{aligned}
$$

In this sequence, $\boldsymbol{\phi}^{(S)}$ depends on $\boldsymbol{\phi}^{(0)}, \ldots, \boldsymbol{\phi}^{(S-1)}$ only through $\boldsymbol{\phi}^{(S-1)}$, i.e. $\boldsymbol{\phi}^{(S)}$ is conditionally independent

of $\boldsymbol{\phi}^{(0)}, \ldots, \boldsymbol{\phi}^{(S-2)}$ given $\boldsymbol{\phi}^{(S-1)}$. This is called the Markov property, and so the sequence is called a Markov chain.

Or, each $\phi_i^{(s+1)}$ conditioning on the $\phi_i^{(s)}$ of the previous Gibbs sampling $\boldsymbol{\phi}^{(s)}$ *and* the new samples $\phi_i^{(s+1)}$ as they are received. Then $\boldsymbol{\phi}^{(s+1)} = \begin{pmatrix} \phi_1^{(s+1)} & \cdots & \phi_n^{(s+1)} \end{pmatrix}$, the $s+1$ th sample is only conditionally dependent on the s th sample. Hence the term "Markov Chain Monte Carlo".

6.5.1 Property

1. The sampling distribution of $\boldsymbol{\phi}^{(s)}$ approaches the target distribution as $s \to \infty$, no matter what the starting value $\boldsymbol{\phi}^{(0)}$ is.
$$\Pr(\boldsymbol{\phi}^{(s)} \in A) \to \int_A p(\boldsymbol{\phi})\, d\boldsymbol{\phi} \quad \text{as } s \to \infty.$$
2. We can approximate $E[g(\phi)]$ (g is function of interest) with the sample average of $\{g(\boldsymbol{\phi}^{(1)}), \ldots, g(\boldsymbol{\phi}^{(S)})\}$, just as in Monte Carlo approximation:
$$\frac{1}{S}\sum_{s=1}^{S} g(\phi^{(s)}) \to \mathrm{E}[g(\phi)] = \int g(\phi)p(\phi)d\phi \quad \text{as } S \to \infty.$$
   - Such approximation: **Markov chain Monte Carlo (MCMC)** approximation.
   - Such procedure: an MCMC algorithm

6.5.2 Distinguishing parameter estimation from posterior approximation

Distinguish the part of the data analysis which is *statistical* from that which is *numerical approximation*

The necessary ingredients of a Bayesian data analysis:

1. *Model specification*: a collection of probability distributions $p(\boldsymbol{y}|\phi), \phi \in \Phi$ which should represent the sampling distribution of your data for some value of $\phi \in \Phi$;
2. *Prior specification*: a probability distribution $p(\phi)$, ideally representing someone's prior information about which parameter values are likely to describe the sampling distribution. The posterior $p(\phi|\boldsymbol{y})$ is completely determined:
$$p(\phi|\boldsymbol{y}) = \frac{p(\phi)p(\boldsymbol{y}|\phi)}{p(\boldsymbol{y})} = \frac{p(\phi)p(\boldsymbol{y}|\phi)}{\int p(\phi)p(\boldsymbol{y}|\phi)d\phi}$$
3. *Posterior summary*: a description of the posterior distribution $p(\phi|\boldsymbol{y})$, done in terms of particular quantities of interest such as posterior means, medians, modes, predictive probabilities, and confidence regions.

Monte Carlo samples from $p(\phi|\boldsymbol{y})$: a way to "see" $p(\phi|\boldsymbol{y})$. Thus, MC and MCMC sampling algorithms

- are not models
- they do not generate "more information" than is in $\boldsymbol{y}$ and $p(\phi)$,
- they are simply "ways of looking at" $p(\phi|\boldsymbol{y})$.

Thus:

- **"Estimation"**: describe how we use $p(\phi|\boldsymbol{y})$ to make inference about ϕ
- **"Approximation"**: describe the use of Monte Carlo procedures to approximate integrals

6.6 Introduction to MCMC diagnostics

For any functions g of interest, the purpose of Monte Carlo or Markov chain Monte Carlo approximation is to obtain a sequence of parameter values $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$ such that

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

1. we want the empirical average of $\{g(\phi^{(1)}), \ldots, g(\phi^{(S)})\}$ to approximate the expected value of $g(\phi)$ under a target probability distribution $p(\phi)$ (in Bayesian inference, usually the posterior distribution $p(\phi|\theta)$).
2. Thus, need the empirical distribution of the simulated sequence $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$ to look like the target distribution $p(\phi)$.

Monte Carlo and *Markov chain Monte Carlo* are **two ways** of generating such a sequence.

- Monte Carlo simulation: generate independent samples from the target distribution. Independent MC samples automatically create a sequence that is representative of $p(\phi)$: The probability that $\phi^{(s)} \in A$ for any set A is $\int_A p(\phi)d\phi$.
  This is **true** for every $s \in \{1, \ldots, S\}$ and conditionally or unconditionally on the other values in the sequence.
- This is **not true** for MCMC samples, in which case all we are sure of is that

$$\lim_{s\to\infty} \Pr(\phi^{(s)} \in A) = \int_A p(\phi)\, d\phi.$$

6.6.1 Interpretation

In the case of a generic parameter ϕ and target distribution $p(\phi)$, think of the sequence $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$ as the trajectory of a particle ϕ moving around the parameter space.

MCMC integral approximation: the amount of time the particle spends in a given set A is proportional to the target probability $\int_A p(\phi)d\phi$.

Suppose A_1, A_2 and A_3 are disjoint subsets of the parameter space, with $Pr(A_2) < Pr(A_1) \approx Pr(A_3)$. In terms of the integral approximation, we want the particle to spend little time in A_2, and about the same amount of time in A_1 as in A_3. It is possible that we would accidentally start our Markov chain in A_2 because $p(\phi)$ is unknown. Thus, it is critical that the number of iterations S is large enough so that the particle has a chance to

(1) move out of A_2 and into higher probability regions $\rightarrow$ chain has achieved stationarity or converged
Thus, one thing to check for is *stationarity*, or that samples taken in one part of the chain have a similar distribution to samples taken in other parts.
- For semiconjugate prior, stationarity can be achieved quickly
- For some highly parameterized models, it takes a long time to be stational

(2) move between A_1 and A_3, and any other sets of high probability

6.6.2 About Item 2

Item 2 above (called the speed of *mixing*): how quickly the particle moves around the parameter space.

- An independent MC sampler has perfect mixing: It has **zero autocorrelation** and can jump between different regions of the parameter space in one step.

- An MCMC sampler might have poor mixing, take a long time between jumps to different parts of the parameter space and have a **high degree of autocorrelation**.

How does the correlation of the MCMC samples affect posterior approximation?

Suppose: Approximate the integral $E[\phi] = \int \phi p(\phi) d\phi = \phi_0$ using the empirical distribution of $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$.

1. ϕ-values: independent MC samples from $p(\phi)$, then the variance of $\bar{\phi} = \sum \phi^{(s)}/S$ is

$$\text{Var}_{\text{MC}}[\bar{\phi}] = \text{E}[(\bar{\phi} - \phi_0)^2] = \frac{\text{Var}[\phi]}{S}, \text{where Var}[\phi] = \int \phi^2 p(\phi) d\phi - \phi_0^2$$

Recall from Chapter 4 that the square root of $\text{Var}_{\text{MC}}[\bar{\phi}]$ is the Monte Carlo standard error, and is a measure of how well we expect $\bar{\phi}$ to approximate the integral $\int \phi p(\phi) d\phi$. If we were to rerun the MC approximation procedure many times, perhaps with different starting values or random number generators, we expect that ϕ_0, the true value of the integral, would be contained within the interval $\bar{\phi} \pm 2\sqrt{\text{Var}_{\text{MC}}[\bar{\phi}]}$ for roughly 95% of the MC approximations. We can make this as small as we want by generating more MC samples.

2. MCMC (such as the Gibbs sampler)
Assuming stationarity has been achieved, the expected squared difference from the MCMC integral approximation $\bar{\phi}$ to the target $\phi_0 = \int \phi p(\phi) d\phi$ is the MCMC variance, and is given by

$$\begin{aligned}
\text{Var}_{\text{MCMC}}[\bar{\phi}] &= \text{E}[(\bar{\phi} - \phi_0)^2] = \text{E}[\{\frac{1}{S}\sum(\phi^{(s)} - \phi_0)\}^2] \\
&= \frac{1}{S^2}\text{E}[\sum_{s=1}^{S}(\phi^{(s)} - \phi_0)^2 + \sum_{s \neq t}(\phi^{(s)} - \phi_0)(\phi^{(t)} - \phi_0)] = \frac{1}{S^2}\sum_{s=1}^{S}\text{E}[(\phi^{(s)} - \phi_0)^2] + \frac{1}{S^2}\sum_{s \neq t}\text{E}[(\phi^{(s)} - \phi_0)(\phi^{(t)} - \phi_0)] \\
&= \text{Var}_{\text{MC}}[\bar{\phi}] + \frac{1}{S^2}\sum_{s \neq t}\text{E}[(\phi^{(s)} - \phi_0)(\phi^{(t)} - \phi_0)].
\end{aligned}$$

So the MCMC variance is equal to the *MC variance* plus *a term that depends on the correlation of samples within the Markov chain*. This term is generally positive and so the MCMC variance is higher than the MC variance, meaning that we expect the MCMC approximation to be further away from ϕ_0 than the MC approximation is. The higher the autocorrelation in the chain, the larger the MCMC variance and the worse the approximation is.

How much correlation there is in the chain?

Compute the sample autocorrelation function:

For a generic sequence of numbers $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$, the lag-$t$ autocorrelation function estimates the correlation between elements of the sequence that are t steps apart:

$$\text{acf}_t(\phi) = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\phi_s - \bar{\phi})(\phi_{s+t} - \bar{\phi})}{\frac{1}{S-1}\sum_{s=1}^{S}(\phi_s - \bar{\phi})^2}$$

How much sample?

S_{eff}, the number of independent MC samples necessary to give the same precision as the MCMC samples:

$$\text{Var}_{\text{MCMC}}[\bar{\phi}] = \frac{\text{Var}[\phi]}{S_{\text{eff}}}$$

The higher the autocorrelation, the more MCMC samples we need to attain a given level of precision for our approximation.

7 The multivariate normal model

- Before: univariate models → models for a single measurement on each member of a sample of individuals or each run of a repeated experiment
- Now: datasets are frequently multivariate, having multiple measurements for each individual or experiment → the most useful model for multivariate data, the multivariate normal model

7.1 The multivariate normal density

7.1.1 Example: reading comprehension

We make the step to two variables by example. Consider a sample ($n = 22$) of children who are given reading comprehension tests before and after receiving a method. We can denote these *two* variables for student i as a vector $\mathbf{Y}_i$, where $Y_{i,1}$ is the before score, and $Y_{i,2}$ is the after score:

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix}, \quad \mathrm{E}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{E}[Y_{i,1}] \\ \mathrm{E}[Y_{i,2}] \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \boldsymbol{\theta}$$

$$\Sigma = \mathrm{Cov}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{E}[Y_1^2] - \mathrm{E}[Y_1]^2 & \mathrm{E}[Y_1 Y_2] - \mathrm{E}[Y_1]\,\mathrm{E}[Y_2] \\ \mathrm{E}[Y_1 Y_2] - \mathrm{E}[Y_1]\,\mathrm{E}[Y_2] & \mathrm{E}[Y_2^2] - \mathrm{E}[Y_2]^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

$$\text{Correlation coefficient}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)} = \frac{\sigma_{1,2}}{\sigma_1 \sigma_2} = \frac{\sigma_{1,2}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

7.1.2 Multivariate normal density

- a univariate normal model: (θ, σ^2), or equivalently its first two moments $E[Y] = \theta, E[Y^2] = \sigma^2 + \theta^2$
- a multivariate normal model (in 7.1.1): use the first-order moments $E[Y_1], E[Y_2]$ and the second-order moments $E[Y_1^2], E[Y_1 Y_2], E[Y_2^2]$

If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$ (a p-dimensional data vector $\boldsymbol{Y}$ has a multivariate normal distribution) where

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

then

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\theta}, \Sigma) &= (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\theta}) \right) \\ &= \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\theta}) \right) \end{aligned}$$

where $|\boldsymbol{A}|$ is the determinant of $\boldsymbol{A}$ measureing how "big" A is; $\boldsymbol{b}^T \boldsymbol{A}$ is equal to the $1 \times p$ vector

$(\sum_{j=1}^p b_j a_{j,1}, \ldots, \sum_{j=1}^p b_j a_{j,p})$, and $\boldsymbol{b}^T \boldsymbol{A} \boldsymbol{b}$ is the number $\sum_{j=1}^p \sum_{k=1}^p b_j b_k a_{j,k}$.

An interesting feature of the multivariate normal distribution: the marginal distribution of each variable Y_j is a univariate normal distribution, with mean θ_j and variance σ_j^2.

7.2 A semiconjugate prior distribution for the mean

A full closed-form solution for the posterior → too complicated. → Compute conjugate priors and posteriors for the full conditional distributions of the $\boldsymbol{\theta}$ and Σ separately, and use Gibbs sampling to easily estimate the joint posterior distribution. Our goal is to obtain the full conditional distribution $\{\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\}$

First, calculating $\{\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\}$:

Analogous to the univariate case, for the multivariate normal distribution, a conjugate prior to the population mean is a multivariate normal: Let $\boldsymbol{\mu}_0$ be the prior mean, and Λ_0 be the covariance matrix of $\boldsymbol{\mu}_0$. Then we belief $\boldsymbol{\theta} \mid \Sigma \sim \mathcal{N}(\boldsymbol{\mu}_0, \Lambda_0)$. This prior:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \Sigma) &= (2\pi)^{-p/2}|\Lambda_0|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T A_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)\} \\
&= (2\pi)^{-p/2}|\Lambda_0|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta} - \boldsymbol{\mu}_0^T\Lambda_0^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T\Lambda_0^{-1}\boldsymbol{\mu}_0)\right] \\
&= (2\pi)^{-p/2}|\Lambda_0|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_0^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right) \qquad (7.1) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right) \\
&= \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}_0\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{b}_0\right),
\end{aligned}
$$

where $\mathbf{A}_0 = \Lambda_0^{-1}$, i.e. the precision matrix (which echoes the univariate case) and $\boldsymbol{b}_0 = \Lambda_0^{-1}\boldsymbol{\mu}_0 = \mathbf{A}_0\boldsymbol{\mu}_0$

We will see this simplified form show up when working with the sampling models and posterior distribution.

Let's first look at the sampling model/likelihood function. The sampling model is that $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \mid \boldsymbol{\theta}, \Sigma\}$ are i.i.d. multivariate $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$. Thus, the joint sampling density of the observed vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is:

$$\begin{aligned} p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | \boldsymbol{\theta}, \Sigma) &= \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta})\} \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta})\} \\ &\propto \exp\{-\frac{1}{2} \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} + \boldsymbol{\theta}^T \boldsymbol{b}_1\} \end{aligned}$$

where $\mathbf{A}_1 = n\Sigma^{-1}$ and $\boldsymbol{b}_1 = n\Sigma^{-1}\bar{\boldsymbol{y}}$. $\bar{\boldsymbol{y}}$ is the vector of sample averages for each variable:

$$\bar{\boldsymbol{y}} = (\frac{1}{n} \sum_{i=1}^{n} y_{i,1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} y_{i,p})^T.$$

Thus, the posterior for $\boldsymbol{\theta}$ is:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma) &= p(\boldsymbol{\theta} \mid \Sigma) p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\theta}, \Sigma) / p(\mathbf{y}_1, \ldots, \boldsymbol{b}my_n \mid \Sigma) \propto p(\boldsymbol{\theta} \mid \Sigma) p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\theta}, \Sigma) \\ &\propto \exp\left(-\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{A}_0 \boldsymbol{\theta} + \boldsymbol{\theta}^T \boldsymbol{b}_0\right) \times \exp(-\frac{1}{2} \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} + \boldsymbol{\theta}^T \boldsymbol{b}_1) \\ &= \exp\left(-\frac{1}{2} \boldsymbol{\theta}^T \mathbf{A}_n \boldsymbol{\theta} + \boldsymbol{\theta}^T \boldsymbol{b}_n\right) \end{aligned}$$

where we have combined terms such that $\mathbf{A}_n = \mathbf{A}_0 + \mathbf{A}_1 = \Lambda_0^{-1} + n\Sigma^{-1}$ and $\boldsymbol{b}_n = \boldsymbol{b}_0 + \boldsymbol{b}_1 = \Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}}$.

So if $\boldsymbol{\theta} \mid \Sigma \sim \mathcal{N}(\boldsymbol{\mu}_0, \Lambda_0)$, then $\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma \sim \mathcal{N}(\boldsymbol{\mu}_n, \Lambda_n)$. The conditional distribution of $\boldsymbol{\theta}$ therefore must be a multivariate normal distribution with covariance $\mathbf{A}_n^{-1}$ and mean $\mathbf{A}_n^{-1}\boldsymbol{b}_n$,

Thus, like the univariate case:

- $\text{Cov}(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma) = \Lambda_n = (A_0^{-1} + n\Sigma^{-1})^{-1}$
  a combination of prior and posterior precision
- $\mathbb{E}(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma) = \boldsymbol{\mu}_n = (A_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}})$
  a weighted average of the prior estimate of the mean and the sample mean.
- $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma)$ = multivariate normal $(\boldsymbol{\mu}_n, \Lambda_n)$

7.3 The inverse-Wishart distribution

The (semi-)conjugate prior and posterior distribution for the mean $\boldsymbol{\theta}$ ✓. Now: the covariance matrix Σ

- For the univariate case, a semi-conjugate prior distribution for the variance σ^2 was the inverse-Gamma distribution ($\sigma^2 \sim IG$; $1/\sigma^2 \sim Ga$)
- For the multivariate case, a semi-conjugate prior distribution for the covariance matrix Σ is the inverse of the multivariate analog of the Gamma distribution, known as a Wishart distribution ($\Sigma \sim IW$; $\Sigma^{-1} \sim W$)

The inverse-Wishart distribution is used to make sure the variance-covariance matrix Σ is:

- *positive definite*, that is: $\boldsymbol{x}'\Sigma\boldsymbol{x} > 0$ for all vectors $\boldsymbol{x}$
- *symmetric*, that is: $\sigma_{j,k} = \sigma_{k,j}$

so that it can be a valid prior distribution.

7.3.1 Empirical covariance matrices

Denote $\boldsymbol{z}_i$ as a $p \times 1$ vector, then $\boldsymbol{z}_i \boldsymbol{z}_i^T$ is a $p \times p$ matrix as follow:

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

If $\boldsymbol{z}_i$ are samples from a population with zero mean, we can think of the matrix $\boldsymbol{z}_i \boldsymbol{z}_i^T / n$ as the contribution of vector $\boldsymbol{z}_i$ to the estimate of the covariance matrix of all of the observations. The *sum of squares matrix* of a collection of n p-dimensional multivariate vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ is: $\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^T = \mathbf{Z}^T \mathbf{Z}$, where $\mathbf{Z}$ is the $n \times p$ matrix whose i-th row is $\boldsymbol{z}_i^T$. Thus, if we divide the sum of squares matrix by n, we get a *sample covariance matrix*, an unbiased estimator of the population covariance matrix.

If $n > p$ and the $\boldsymbol{z}_i$ are linearly independent, then $\mathbf{Z}^T \mathbf{Z}$ will be positive definite and symmetric.

To summarize, generating samples from a Wishart distribution is analogous to sampling a set of variables from a multivariate normal distribution and calculating the empirical covariance matrix of the samples. More specifically, with parameters $\nu_0 \in \mathbb{Z}^+$ and $\mathbf{S}_0$ (a $p \times p$ covariance matrix),

1. Sample $z_1, \ldots, z_{\nu_0} \sim$ (i.i.d.)$\mathcal{N}(\mathbf{0}, \mathbf{S}_0^{-1})$
2. Calculate $\boldsymbol{Z}^T \boldsymbol{Z} = \sum_{i=1}^{\nu_0} z_i z_i^T$
3. Repeat this procedure over and over again, generating matrices (sum of squares matrices) $\boldsymbol{Z}_1^T \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_S^T \boldsymbol{Z}$

Then $\boldsymbol{Z}_1^T \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_S^T \boldsymbol{Z} \sim \text{Wishart}(\nu_0, \mathbf{S}_0^{-1})$.

Some properties of samples from the Wishart with parameters $(\nu_0, \mathbf{S}_0^{-1})$:

- If $\nu_0 > p$ then $\boldsymbol{Z}^T\boldsymbol{Z}$ is positive definite
- $\boldsymbol{Z}^T\boldsymbol{Z}$ is symmetric
- $\mathbb{E}(\boldsymbol{Z}^T\boldsymbol{Z}) = \nu_0\mathbf{S}_0^{-1}$

Recall the univariate case, we have:

- Wishart distribution is a semi-conjugate prior distribution for the precision matrix Σ^{-1}
- inverse-Wishart distribution is a semi-conjugate prior distribution for the covariance matrix Σ.

To sample a covariance matrix Σ from an inverse-Wishart distribution, we first follow the above step and set $\Sigma^{(s)} = (\boldsymbol{Z}^{(s)T}\boldsymbol{Z}^{(s)})^{-1}$ (actually $\Sigma = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$)

Note that $\mathbb{E}(\Sigma^{-1}) = \nu_0\mathbf{S}_0^{-1}$

$$\mathbb{E}(\Sigma) = \frac{1}{\nu_0 - p - 1}(\mathbf{S}_0)^{-1} = \frac{1}{\nu_0 - p - 1}\mathbf{S}_0 \text{(so not exactly the inverse of } \mathbf{S}_0^{-1})$$

Specifying parameters

If we have a prior expectation of a covariance matrix Σ_0, then we can center our prior around the empirical covariance matrix in two suggested ways:

- Set ν_0 large and set $\mathbf{S}_0 = (\nu_0 - p - 1)\Sigma_0$, such that $\mathbb{E}(\Sigma) = \frac{\nu_0-p-1}{\nu_0-p-1}\Sigma_0 = \Sigma_0$ and (due to large ν_0) the prior is fairly concentrated around Σ_0
- Set $\nu_0 = p + 2$ and let $\mathbf{S}_0 = \Sigma_0$, such that $\mathbb{E}(\Sigma) = \frac{1}{p+2-p-1}\Sigma_0 = \Sigma_0$ but only loosely centered around Σ_0 (due to fairly small ν_0)

7.3.2 Full conditional distribution of the covariance matrix, $\Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\theta}$

The inverse-Wishart$(\nu_0, \mathbf{S}_0^{-1})$ density is given by

$$p(\Sigma) = \left[2^{\nu_0 p/2} \pi^{\binom{p}{2}/2} |\mathbf{S}_0|^{-\nu_0/2} \prod_{j=1}^{p} \Gamma([\nu_0 + 1 - j]/2) \right]^{-1} \times |\Sigma|^{-(\nu_0+p+1)/2} \times \exp\{-\text{tr}(\mathbf{S}_0\Sigma^{-1})/2\}.$$

But all we need to know is if $\Sigma \sim$ inverse-Wishart$(\nu_0, \mathbf{S}_0^{-1})$,

$$p(\Sigma) \propto |\Sigma|^{-(\nu_0+p+1)/2} \times \exp\left(-\text{tr}(\mathbf{S}_0\Sigma^{-1})/2\right)$$

where "tr" stands for trace and for a square $p \times p$ matrix $\mathbf{A}$, $\text{tr}(\mathbf{A}) = \sum_{j=1}^{p} a_{j,j}$, sum of the diagonal elements. Besides, $tr(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{A}) = \sum_{k=1}^{K} \boldsymbol{b}_k^T \boldsymbol{A} \boldsymbol{b}_k$ where $\boldsymbol{B}$ is the matrix whose k-th row is $\boldsymbol{b}_k^T$.

The sampling model (likelihood):

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | \boldsymbol{\theta}, \Sigma) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) \right)$$

Using some linear algebra,

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) = \text{tr}\left(\left(\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T \right) \Sigma^{-1} \right) = \text{tr}\left(\mathbf{S}_\theta \Sigma^{-1}\right)$$

where $\mathbf{S}_\theta = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T$ is the *residual sum of squares matrix* for the vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$.

Recall that, since inverse-Wishart matrices involve sampling from a normal distribution with mean $\mathbf{0}$, indeed $\mathbf{S}_0$ can be treated as a *residual* covariance matrix, given that $\mathbb{E}(\boldsymbol{y}_i - \boldsymbol{\theta}) = \mathbf{0}$. To obtain the residual sum of squares matrix, you calculate the sum of squares for the residual vectors $\boldsymbol{y}_i - \boldsymbol{\theta}$; Conditional on $\boldsymbol{\theta}$, $\frac{1}{n}\mathbf{S}_\theta$ provides an unbiased estimate of the true covariance matrix $\text{Cov}[\boldsymbol{Y}]$.

Thus

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\theta}, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2} \exp\left(-\text{tr}(\mathbf{S}_\theta \Sigma^{-1})/2\right)$$

Now we can calculate the full conditional distribution of Σ:

$$\begin{aligned}
p(\Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\theta}) &\propto p(\Sigma) \times p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\theta}, \Sigma) \\
&\propto \left[|\Sigma|^{-(\nu_0+p+1)/2} \times \exp\left(-\text{tr}(\mathbf{S}_0\Sigma^{-1})/2\right)\right] \times \left[|\Sigma|^{-n/2} \exp\left(-\text{tr}(\mathbf{S}_\theta\Sigma^{-1})/2\right)\right] \\
&= |\Sigma|^{-(\nu_0+n+p+1)/2} \exp\left\{-\text{tr}([\mathbf{S}_0 + \mathbf{S}_\theta]\,\Sigma^{-1})/2\right\} \\
&\propto \text{dinverse-Wishart}\left(\nu_0 + n, [\mathbf{S}_0 + \mathbf{S}_\theta]^{-1}\right) \\
&= \text{dinverse-Wishart}\left(\nu_n, \mathbf{S}_n^{-1}\right)
\end{aligned}$$

where $\nu_n = \nu_0 + n$ and $\mathbf{S}_n = \mathbf{S}_0 + \mathbf{S}_\theta$.

Like the univariate case, the full conditional distribution on Σ is dependent on (1)$\nu_n = \nu_0 + n$, a "posterior sample size", which is a sum of the prior sample size ν_0 and the data sample size n, and (2)$\mathbf{S}_n = \mathbf{S}_0 + \mathbf{S}_\theta$, the "prior" residual sum of squares + the residual sum of squares from the data(the empirical sum of squares).

Thus we have

$$\{\Sigma|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\theta}\} \sim \text{inverse- Wishart}(\nu_0 + n, [\mathbf{S}_0 + \mathbf{S}_\theta]^{-1}) = \text{inverse- Wishart}(\nu_n, \mathbf{S}_n^{-1})$$

Finally, notice that the conditional expectation of the covariance matrix is a weighted average of the prior expectation $\frac{1}{\nu_0-p-1}\mathbf{S}_0$ and the unbiased estimator $\frac{1}{n}\mathbf{S}_\theta$:

$$\begin{aligned}\mathbb{E}(\Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\theta}) &= \frac{1}{\nu_n - p - 1}(\mathbf{S}_n) = \frac{1}{\nu_0 + n - p - 1}(\mathbf{S}_0 + \mathbf{S}_\theta) \\ &= \frac{\nu_0 - p - 1}{\nu_0 + n - p - 1}\frac{1}{\nu_0 - p - 1}\mathbf{S}_0 + \frac{n}{\nu_0 + n - p - 1}\frac{1}{n}\mathbf{S}_\theta.\end{aligned}$$

7.4 Summary of inference with the multivariate normal

Like in Chapter 5, here we summarize the moving parts of inference with the multivariate normal sampling model. There are four prior parameters (note some are matrices):

(Semiconjugate) prior

- $\mathbf{S}_0$ for the inverse-Wishart
  - *related to* the prior estimate of the covariance between the variables
  - Only *related* to because as mentioned above, there are some guidelines for what to use for ν_0 and $\mathbf{S}_0$ such that the prior distribution is centered around Σ_0, the **true* prior estimate of the covariance matrix you are looking for.
- ν_0 for the inverse-Wishart
  - a "prior sample size" from which the initial estimate of the *variance* is observed
- $\boldsymbol{\mu}_0$ for the multivariate normal
  - an initial estimate for the population mean
- Λ_0 for the multivariate normal

- the covariance (i.e. uncertainty) of the initial estimate for the population mean

In other words, we have these semiconjugate priors (or multivariate normal-inverse-Wishart prior):

$$\{\Sigma\} \sim \text{inverse-Wishart}(\nu_0, \mathbf{S}_0^{-1})$$
$$\{\boldsymbol{\theta} \mid \Sigma\} \sim \text{multivariate normal}(\boldsymbol{\mu}_0, \Lambda_0)$$

Similarly to the univariate case, the estimate of the covariance matrix for the inverse-Wishart prior is decoupled from the estimate of the covariance of the mean vector in the multivariate normal prior, although it's common to set these the same.

Note that this is somewhat *different* than the univariate case; since there were no covariances to worry about, what was decoupled was "prior sample sizes" from which the prior variance and prior mean are observed. Like here, it was also common to set these the same.

Posterior The updated parameters are

- $\mathbf{S}_n = \mathbf{S}_0 + \mathbf{S}_\theta$, where $\mathbf{S}_\theta$ is the residual sum of squares matrix
- $\nu_n = \nu_0 + n$
- $\boldsymbol{\mu}_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}}) = \Lambda_n(\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}})$
- $\Lambda_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$

The full conditional posterior distributions are:

$$\{\Sigma|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\theta}\} \sim \text{inverse-Wishart}(\nu_n, \mathbf{S}_n^{-1})$$
$$\{\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\} \sim \text{multivariate normal}(\boldsymbol{\mu}_n, \Lambda_n)$$

7.5 Gibbs sampling of the mean and covariance

$$\{\boldsymbol{\theta}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\Sigma\} \sim \text{multivariate normal}(\boldsymbol{\mu}_n,\Lambda_n)$$
$$\{\Sigma|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n,\boldsymbol{\theta}\} \sim \text{inverse-Wishart}(\nu_n,\mathbf{S}_n^{-1}),$$

Knowing these values, we can now perform Gibbs sampling to sample from $p(\boldsymbol{\theta},\Sigma \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y_n})$. Specifically, we start with an estimate of one of the two values-$\Sigma^{(0)}$ for simplicity- and use the following two steps:

1. Sample $\boldsymbol{\theta}^{(s+1)} \sim \mathcal{N}(\boldsymbol{\mu}_n,\Lambda_n)$. This depends on the inverse of the previous $\Sigma^{(s)}$.

   a) compute $\boldsymbol{\mu}_n$ and Λ_n from $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ and $\Sigma^{(s)}$;
   b) sample $\boldsymbol{\theta}^{(s+1)} \sim$ multivariate normal$(\boldsymbol{\mu}_n,\Lambda_n)$.

2. Sample $\Sigma^{(s+1)} \sim$ inverse-Wishart$(\nu_n,\mathbf{S}_n^{-1})$, where the parameters depend on $\boldsymbol{\theta}^{(s+1)}$.

   a) compute $\mathbf{S}_n$ from $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ and $\boldsymbol{\theta}^{(s+1)}$;
   b) sample $\Sigma^{(s+1)} \sim$ inverse-Wishart$(\nu_0+n,\mathbf{S}_n^{-1})$.

7.6 Missing data and imputation

The posterior distribution for $\boldsymbol{\theta}$ and Σ depends on $\prod_{i=1}^{n} p(\boldsymbol{y}_i|\boldsymbol{\theta}, \Sigma)$, but $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}, \Sigma)$ cannot be computed if components of $\boldsymbol{y}_i$ are missing. Either throw away all subjects with incomplete data or impute missing values with a population mean or some other fixed value is incorrect.

How to handle this? A map is: Let $\boldsymbol{O}_i = (O_1, \ldots, O_p)^T$ be a binary vector of zeros and ones (thus $\boldsymbol{O}$ is a $p \times j$ matrix, p is the number of samples and j is the order of variables), and define:

$$O_{i,j} = \begin{cases} 1 & \text{if } Y_{i,j} \text{ is observed} \\ 0 & \text{if } Y_{i,j} \text{ is missing} \end{cases}$$

Assume that missing data are *missing at random*, meaning that $\boldsymbol{O}_i$ and $\boldsymbol{Y}_i$ are statistically independent and that the distribution of $\boldsymbol{O}_i$ does not depend on $\boldsymbol{\theta}$ or Σ. Our sampling probability for data from subject i is $p(\boldsymbol{o}_i)$ multiplied by the marginal probability of the observed variables, after integrating out the missing variables:

$$\begin{aligned} p(\boldsymbol{o}_i, \{y_{i,j} : o_{i,j} = 1\}|\boldsymbol{\theta}, \Sigma) &= p(\boldsymbol{o}_i) \times p(\{y_{i,j} : o_{i,j} = 1\}|\boldsymbol{\theta}, \Sigma) \\ &= p(\boldsymbol{o}_i) \times \int \left\{ p(y_{i,1}, \ldots, y_{i,p}|\boldsymbol{\theta}, \Sigma) \prod_{y_{i,j}:o_{i,j}=0} dy_{i,j} \right\}. \end{aligned}$$

The correct thing is to integrate over the missing data to obtain the marginal probability of the observed data.

But combining marginal densities from subjects having different amounts of information can be notationally awkward. Fortunately, our integration can alternatively be done quite easily using Gibbs sampling.

7.6.1 Gibbs sampling with missing data

We can use Gibbs sampling to estimate the posterior $p(\boldsymbol{\theta}, \Sigma \mid \mathbf{Y})$.

Here, however, we don't have a full dataset $\mathbf{Y}$ (a $n \times p$ matrix); rather, we have an observed dataset $\mathbf{Y}_{\text{obs}}$ and missing values $\mathbf{Y}_{\text{miss}}$. We can still use the Gibbs sample, and the key idea is to also estimate the posterior distribution on $\mathbf{Y}_{\text{miss}}$, which will also help us make more accurate estimates on $\boldsymbol{\theta}$ and Σ. Our goal is to obtain $p(\boldsymbol{\theta}, \Sigma, \mathbf{Y}_{miss} \mid \mathbf{Y}_{obs})$, the posterior distribution of unknown and unobserved quantities.

Given starting values $\Sigma^{(0)}$ and $\mathbf{Y}_{\text{miss}}^{(0)}$ - perhaps the empirical covariance matrix and the unconditional means of the observed sample, we generate $\{\boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)}, \mathbf{Y}_{\text{miss}}^{(s+1)}\}$ from $\{\boldsymbol{\theta}^{(s)}, \Sigma^{(s)}, \mathbf{Y}_{\text{miss}}^{(s)}\}$ by:

1. Sample $\boldsymbol{\theta}^{(s+1)}$ from $p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}^{(s)}, \Sigma^{(s)})$
2. Sample $\Sigma^{(s+1)}$ from $p(\Sigma \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}^{(s)}, \boldsymbol{\theta}^{(s+1)})$
3. Sample $\mathbf{Y}_{\text{miss}}^{(s+1)}$ from $p(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(s)}, \Sigma^{(s+1)})$

For Steps 1 and 2, you simply combine the sampled missing data $\mathbf{Y}_{\text{miss}}^{(s)}$ and the observed data $\mathbf{Y}_{\text{obs}}$for a full dataset $\mathbf{Y}$ and sample from the full conditional distributions like normal.

For Step 3, Noted that once we've sampled a set of missing values, notice that we now have a full "dataset" if we combine our observed values with the newly sampled missing values. This means that we can sample from the full conditional distributions of $\boldsymbol{\theta}$ and Σ normally, and from there, once again sample a new set of $\mathbf{Y}_{\text{miss}}$.

Before this, we need to obtain $p(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(s)}, \Sigma^{(s)})$. Thus:

$$\begin{aligned} p(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}, \Sigma) &\propto p(\mathbf{Y}_{\text{miss}}, \mathbf{Y}_{\text{obs}} \mid \theta, \Sigma) \\ &= \prod_{i=1}^{n} p(y_{i,\text{miss}}, \boldsymbol{y}_{i,\text{obs}} \mid \boldsymbol{\theta}, \Sigma) \\ &\propto \prod_{i=1}^{n} p(y_{i,\text{miss}} \mid \boldsymbol{y}_{i,\text{obs}}, \boldsymbol{\theta}, \Sigma), \end{aligned}$$

Specifically, to sample from the above, we simply sample the missing values of each data point independently. Thus, we need to know the $p(y_{i,\text{miss}} \mid \boldsymbol{y}_{i,\text{obs}}, \boldsymbol{\theta}, \Sigma)$. The following properties can help us:

For a data point $\boldsymbol{y}$ with missing values, let a be the indices of the observed values and b be the indices of the missing values. Then it is shown that sampling $\boldsymbol{y}_{[b]}$ given known observed variables and the parameters $\boldsymbol{\theta}$ and Σ also follows a multivariate normal distribution, but with mean and covariance matrices with dimension $|b|$ that take into account the existing variables:

$$\begin{aligned} \{\boldsymbol{y}_{[b]} | \boldsymbol{y}_{[a]}, \boldsymbol{\theta}, \Sigma\} &\sim \text{multivariate normal}(\boldsymbol{\theta}_{b|a}, \Sigma_{b|a}), \text{ where} \\ \boldsymbol{\theta}_{b|a} &= \boldsymbol{\theta}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\boldsymbol{y}_{[a]} - \boldsymbol{\theta}_{[a]}) \\ \Sigma_{b|a} &= \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}. \end{aligned}$$

where $\boldsymbol{\theta}_{[b]}$ refers to the elements of $\boldsymbol{\theta}$ corresponding to the indices in $\boldsymbol{b}$, and $\Sigma_{[a,b]}$ refers to the matrix made up of the elements that are in rows $\boldsymbol{a}$ and columns $\boldsymbol{b}$ of Σ.

- Intuitively, the mean of the multivariate normal distribution on the missing values *given* some observed values starts with *the unconditional mean of the observed values*, plus or minus *some offset that depends on the observed values and the correlations between the observed and missing values*.
  For example, if it is known that a datapoint's observed values are quite high relative to the mean $(\boldsymbol{y}_a - \boldsymbol{\theta}_a)$, and that there is a positive correlation between observed values and missing values, we would expect the missing values to generally be higher as well.
- Intuitively, the covariance matrix of the conditional distribution on the missing values starts with the unconditional covariance, but notice the minus sign; since the covariance matrix is positive definite, knowing about some observed variables will decrease our uncertainty about the missing values.

Therefore, we can use the Gibbs sampler to achieve our goal.

Prediction and regression

$$\mathrm{E}[\boldsymbol{y}_{[b]}|\boldsymbol{\theta}, \Sigma, \boldsymbol{y}_{[a]}] = \boldsymbol{\theta}_{[b]} + \boldsymbol{\beta}_{b|a}^T(\boldsymbol{y}_{[a]} - \boldsymbol{\theta}_{[a]})$$

where $\boldsymbol{\beta}_{b|a}^T = \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}$. Since this takes the form of a linear regression model, we call the value of $\boldsymbol{\beta}_{b|a}$ the regression coefficient for $\boldsymbol{y}_{[b]}$ given $\boldsymbol{y}_{[a]}$ based on Σ.

7.7 Discussion and further references

The multivariate normal model can be justified as a sampling model for reasons analogous to those for the univariate normal model (see Section 5.7): It is characterized by independence between the sample mean and sample variance (Rao, 1958), it is a maximum entropy distribution and it provides consistent estimation of the population mean and variance, even if the population is not multivariate normal.

8 Group comparisons and hierarchical modeling

A common task in data analysis is to compare summary statistics for two or more groups. In this chapter we cover the Bayesian approach to doing this. Specifically, we discuss models for the comparison of means across groups.

Parameterize the two population means by their average and their difference. $\rightarrow$ extended to the multigroup case, where the average group mean and the differences across group means are described by a normal sampling model. $\rightarrow$ This model, together with a normal sampling model for variability among units within a group, make up a hierarchical normal model that describes both within-group and between-group variability.

8.1 Comparing two groups

(A commonly taught data analysis procedure) A standard method in (frequentist) introductory statistics for comparing the means of two populations is to compute the t-statistic of the observed mean difference and obtain the two-sided p-value.

$$t(\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}}, \quad s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$$

Then, if $p < 0.05$ (or any other significance level), we reject the model that the two groups have the same distribution, conclude that $\theta_1 \neq \theta_2$, and use the estimates $\hat{\theta}_1 = \bar{y}_1$ and $\hat{\theta}_2 = \bar{y}_2$ (i.e. the ML estimators for θ_1 and θ_2 independently). Otherwise, we accept the model that the two groups have the same distribution, conclude that $\theta_1 = \theta_2$, and let $\hat{\theta}_1 = \hat{\theta}_2 = (\sum y_{i,1} + \sum y_{i,2})/(n_1 + n_2)$ be the pooled mean of the two groups.

Sometimes this paradigm doesn't make too much sense. Consider borderline in which our p-value is close to 0.05. It seems like a technicality to treat the means as completely different if $p = 0.051$, and completely the same if $p = 0.049$- a difference that could hypothetically be observed by simply sampling one more data point.

The Bayesian approach is to treat the two populations as being sampled from a common mean θ plus some difference δ, where we estimate both θ and δ. Then the observed difference δ can vary continuously. Specifically, our sampling model for a value from either group is

$$Y_{i,1} = \mu + \delta + \epsilon_{i,1}, \quad Y_{i,2} = \mu - \delta + \epsilon_{i,2}$$

where we assume values from both groups have a common variance $\epsilon_{i,j} \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$. Besides, $\theta_1 = \mu + \delta$ and $\theta_2 = \mu - \delta$, δ represents half the population difference in means, as $(\theta_1 - \theta_2)/2 = \delta$, and μ represents the pooled average, as $(\theta_1 + \theta_2)/2 = \mu$.

8.1.1 Prior and posterior distributions

Prior The joint prior for all three parameters of our model μ, δ, σ^2 is unsurprising. We treat the parameters as independent, so $p(\mu, \delta, \sigma^2) = p(\mu)p(\delta)p(\sigma^2)$ where

- $\mu \sim \mathcal{N}(\mu_0, \gamma_0^2)$
- $\delta \sim \mathcal{N}(\delta_0, \tau_0^2)$
- $\sigma^2 \sim$ inverse-gamma$(\nu_0/2, \sigma_0^2\nu_0/2)$

Notice that we specify prior distributions for the common mean and variance, but we also express an estimate δ (and certainty of the estimate) for the difference between the group means.

Posterior

Then the full conditional distributions of the parameters are

- $\mu \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \delta, \sigma^2 \sim \mathcal{N}(\mu_n, \gamma_n^2)$
  - $\mu_n = \gamma_n^2 \times \left[\mu_0/\gamma_0^2 + \sum_{i=1}^{n_1}(y_{i,1} - \delta)/\sigma^2 + \sum_{i=1}^{n_2}(y_{i,2} + \delta)/\sigma^2\right]$
  - $\gamma_n^2 = \left[1/\gamma_0^2 + (n_1 + n_2)/\sigma^2\right]^{-1}$
- $\delta \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \mu, \sigma^2 \sim \mathcal{N}(\delta_n, \tau_n^2)$
  - $\delta_n = \tau_n^2 \times \left[\delta_0/\tau_0^2 + \sum_{i=1}^{n_1}(y_{i,1} - \mu)/\sigma^2 - \sum_{i=1}^{n_2}(y_{i,2} - \mu)/\sigma^2\right]$
  - $\tau_n^2 = \left[1/\tau_0^2 + (n_1 + n_2)/\sigma^2\right]^{-1}$
- $\sigma^2 \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \mu, \delta \sim \text{inverse-gamma}(\nu_n/2, \sigma_n^2\nu_n/2)$
  - $\nu_n = \nu_0 + n_1 + n_2$
  - $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \sum_{i=1}^{n_1}(y_{i,1} - [\mu + \delta])^2 + \sum_{i=1}^{n_2}(y_{i,2} - [\mu - \delta])^2$

8.2 Comparing multiple groups

Let's extend this to a > 2 group case. Assume for our example above that we have many schools, which we assume are samples from a population of schools. So our dataset is *hierarchical* or *multilevel* since there are samples of schools and within each school samples of students, i.e., a hierarchy of nested populations.

Of course, the simplest type of multilevel data has 2 levels, in which one level consists of groups and the other consists of units within groups. In this case we denote $y_{i,j}$ as the data on the ith unit in group j.

8.2.1 Exchangeability and hierarchical models

If exchangeability holds for all values of n, then de Finetti's theorem says that an equivalent formulation of our information is that

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

In other words, the random variables can be thought of as independent samples from a population described by some fixed but unknown population feature ϕ. In the normal model, for example, we take $\phi = \{\theta, \sigma^2\}$ and model the data as conditionally i.i.d. normal(θ, σ^2).

Now consider a model describing our information about a hierarchical data $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\}$, where $\boldsymbol{Y}_j = \{Y_{1,j}, \ldots, Y_{n_j,j}\}$.

What properties should a model $p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ have?

First consider $p(\boldsymbol{y}_j) = p(y_{1,j}, \ldots, y_{n_j,j})$, the marginal probability density of data from a single group j. We can treat $Y_{1,j}, \ldots, Y_{n_j,j}$ as exchangeable. If group j is large compared to the sample size n_j , then we can model the data within group j as conditionally i.i.d. given some group-specific parameter ϕ_j, which we call the *within-group sampling variability*:

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

Further, if we have many groups with parameters ϕ_j that we assume are sampled from a population of *groups* we can again use de Finetti's theorem to treat the group means ϕ_j as conditionally i.i.d. given another

parameter, which we call the *between-group sampling variability*:

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

But how should we represent our information about $\phi_1, \ldots, \phi_m$? If the groups themselves are samples from some larger population of groups, then exchangeability of the group-specific parameters might be appropriate. Applying de Finetti's theorem a second time gives

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

for some sampling model $p(\phi \mid \psi)$ and an unknown parameter ψ. Then we simply need a prior distribution on the parameter for the group parameters (a "hyperparameter") ψ:

$$\psi \sim p(\psi)$$

Note that we can extend this hierarchy arbitrarily:

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

$p(y|\phi)$ and $p(\phi|\psi)$ both represent sampling variability among populations of objects. In contrast, $p(\psi)$ represents information about a single fixed but unknown quantity. For this reason, we refer to $p(y|\phi)$ and $p(\phi|\psi)$ as *sampling distributions*, and are conceptually distinct from $p(\psi)$, which is a *prior distribution*. The data will be only used to estimate the distributions $p(y|\phi)$ and $p(\phi|\psi)$.

8.3 The hierarchical normal model

The hierarchical normal model: treat the data within a group as being normally distributed with some mean θ_j and variance σ^2, and the means among groups to *also* be normally distributed according to some other mean μ and variance τ^2. Specifically, we have

$$\begin{array}{lll} \phi_j = \{\theta_j, \sigma^2\}, & p(y|\phi_j) = \text{normal}(\theta_j, \sigma^2), & Y_{i,j} \mid \phi_j \sim \mathcal{N}(\theta_j, \sigma^2) \quad \text{(within-group model)} \\ \psi = \{\mu, \tau^2\}, & p(\theta_j|\psi) = \text{normal}(\mu, \tau^2) & \theta_j \mid \psi \sim \mathcal{N}(\mu, \tau^2) \quad \text{(between-group model)} \end{array}$$

So for m groups, we have those following unknown parameters: group-specific means $\{\theta_1, \ldots, \theta_m\}$, the within-group variance σ^2 (now we're assuming the data within groups share a common variance σ^2 that doesn't depend on the group j.), and the mean and variance of the group-specific means (μ, τ^2). Notice that there are three fixed parameters for which we need to specify prior distributions: μ, τ^2, and σ^2. For convenience, our priors will be semiconjugate priors:

$$\begin{array}{ll} \sigma^2 \sim \text{inverse-gamma}(\nu_0/2, \sigma_0^2\nu_0/2), & 1/\sigma^2 \sim \text{gamma}(\nu_0/2, \sigma_0^2\nu_0/2) \\ \tau^2 \sim \text{inverse-gamma}(\eta_0/2, \tau_0^2\eta_0/2), & 1/\tau^2 \sim \text{gamma}(\eta_0/2, \tau_0^2\eta_0/2) \\ \mu^2 \sim \mathcal{N}(\mu_0, \gamma_0^2) & \end{array}$$

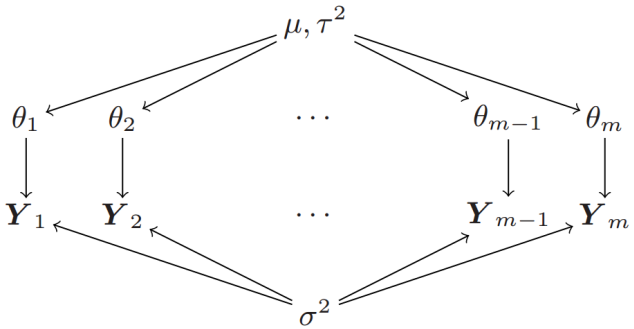


Figure 1: A graphical representation of the basic hierarchical normal model.

8.3.1 Posterior inference

Intuition

We have samples from m groups $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. The unknown quantities: $\{\theta_1, \ldots, \theta_m\}, \sigma^2, (\mu, \tau^2)$. Joint posterior inference for these parameters can be made by constructing a Gibbs sampler that approximates the posterior distribution, and our task is to construct and sample from it:

$$p(\theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2 \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$$

Obtaining the full conditional for a single parameter is fairly straightforward by simply writing the entire joint posterior but then treating the other parameters as constants that can be discarded via proportionality.

To obtain the full joint posterior we will take use key independence assumptions between the parameters of our model. For example, conditionally on $\{\theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2\}$, the random variables $Y_{1,j}, \ldots, Y_{n_j,j}$ are independent with a distribution that depends only on θ_j and σ^2 and not on μ or τ^2. It is helpful to think about

this fact in terms of the diagram in Figure1: The existence of a path from (μ, τ^2) to each $\boldsymbol{Y}_j$ indicates that while (μ, τ^2) provides information about $\boldsymbol{Y}_j$, it only does so indirectly through θ_j, which separates the two quantities in the graph. Thus we have:

$$
\begin{aligned}
& p(\theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2 \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \\
& \propto p(\theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2) \times p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \mid \theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2) && \text{Bayes' rule} \\
& = p(\mu, \tau^2, \sigma^2) \times p(\theta_1, \ldots, \theta_m \mid \mu, \tau^2, \sigma^2) \times p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \mid \theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2) && \text{Chain rule} \\
& = p(\mu)p(\tau^2)p(\sigma^2) \times p(\theta_1, \ldots, \theta_m \mid \mu, \tau^2) \times p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \mid \theta_1, \ldots, \theta_m, \sigma^2) && \text{Indepent} \\
& = p(\mu)p(\tau^2)p(\sigma^2) \times \left[\prod_{j=1}^{m} p(\theta_j \mid \mu, \tau^2)\right] \times \left[\prod_{j=1}^{m} p(\boldsymbol{y}_j \mid \theta_j, \sigma^2)\right] && \text{de Finetti} \\
& = p(\mu)p(\tau^2)p(\sigma^2) \times \left[\prod_{j=1}^{m} p(\theta_j \mid \mu, \tau^2)\right] \times \left[\prod_{j=1}^{m} \left(\prod_{i=1}^{n_j} p(y_{i,j} \mid \theta_j, \sigma^2)\right)\right] && \text{de Finetti 2x}
\end{aligned}
\tag{8.1}
$$

Full conditional distributions

1. **Full conditional distributions of μ and τ^2**
   As a function of μ and τ^2, the term in Eq.8.1 is proportional to

$$
p(\mu)p(\tau^2) \prod_{j=1}^{m} p(\theta_j|\mu, \tau),
$$

   so the full conditional distributions of μ and τ^2 are also proportional to this quantity. Take the full posterior

and discard all terms that don't depend on μ (or τ^2):

$$p(\mu|\theta_1,\ldots,\theta_m,\tau^2,\sigma^2,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m) \propto p(\mu)\prod p(\theta_j|\mu,\tau^2)$$
$$p(\tau^2|\theta_1,\ldots,\theta_m,\mu,\sigma^2,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m) \propto p(\tau^2)\prod p(\theta_j|\mu,\tau^2).$$

which in this case looks exactly like a standard one-sample Normal posterior from Chapter 6, so we borrow that result and replace the relevant variables from our priors. We can do this similarly for the other parameters.

2. **Full conditional distributions of θ_j**
   Collecting the terms in Eq.8.1 that depend on θ_j shows that the full conditional distribution of θ_j must be proportional to

$$p(\theta_j|\mu,\tau^2,\sigma^2,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m) \propto p(\theta_j|\mu,\tau^2)\prod_{i=1}^{n_j} p(y_{i,j}|\theta_j,\sigma^2).$$

   This says that, conditional on $\{\mu,\tau^2,\sigma^2,\boldsymbol{y}_j\}$, θ_j must be conditionally independent of the other θ's as well as independent of the data from groups other than j. Again, it is helpful to refer to Figure 1: While there is a path from each θ_j to every other θ_k, the paths go through (μ,τ^2) or σ^2. We can think of this as meaning that the θ's contribute no information about each other beyond that contained in μ,τ^2 and σ^2.
3. **Full conditional distributions of σ^2**
   σ^2 is conditionally independent of $\{\mu,\tau^2\}$ given $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m,\theta_1,\ldots,\theta_m,\}$. The derivation of the full conditional of σ^2 is similar to that in the one-sample normal model, except now we have information about σ^2 from m separate groups:

$$p(\sigma^2|\theta_1,\ldots,\theta_m,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m) \propto p(\sigma^2)\prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{i,j}|\theta_j,\sigma^2)$$
$$\propto (\sigma^2)^{-\nu_0/2+1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}} (\sigma^2)^{-\sum n_j/2} e^{-\frac{\sum\sum(y_{i,j}-\theta_j)^2}{2\sigma^2}}$$

Note that $\sum\sum(y_{i,j}-\theta_j)^2$ is the sum of squared residuals across all groups, conditional on the within-group means, and so the conditional distribution concentrates probability around a pooled-sample estimate of the variance.

$$\{\mu|\theta_1,\ldots,\theta_m,\tau^2\} \sim \text{normal}\left(\frac{m\bar{\theta}/\tau^2+\mu_0/\gamma_0^2}{m/\tau^2+1/\gamma_0^2}, [m/\tau^2+1/\gamma_0^2]^{-1}\right)$$

.

Quantities

$$\sigma^2 \sim \text{inverse-gamma}(\nu_0/2, \sigma_0^2\nu_0/2)$$
$$\tau^2 \sim \text{inverse-gamma}(\eta_0/2, \tau_0^2\eta_0/2)$$
$$\mu^2 \sim \mathcal{N}(\mu_0, \gamma_0^2)$$

Applying the results of Chapter 6 with the appropriate symbolic replacements, the full conditionals are

$$\{\mu \mid \theta_1, \ldots, \theta_m, \tau^2\} \sim \mathcal{N}\left(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \left[m/\tau^2 + 1/\gamma_0^2\right]^{-1}\right)$$

$$\{\tau^2 \mid \theta_1, \ldots, \theta_m, \mu\} \sim \text{inverse-gamma}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{j=1}^{m}(\theta_j - \mu)^2}{2}\right) \text{ or}$$

$$\{1/\tau^2 | \theta_1, \ldots, \theta_m, \mu\} \sim \text{gamma}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum(\theta_j - \mu)^2}{2}\right)$$

$$\{\theta_j \mid y_{1,j}, \ldots, y_{n_j,j}, \sigma^2\} \sim \mathcal{N}\left(\frac{n_j\bar{y}_j/\sigma^2 + 1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, \left[n_j/\sigma^2 + 1/\tau^2\right]^{-1}\right)$$

$$\{\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{y_1}, \ldots, \boldsymbol{y_n}\} \sim \text{inverse-gamma}\left(\frac{1}{2}\left[\nu_0 + \sum_{j=1}^{m} n_j\right], \frac{1}{2}\left[\nu_0\sigma_0^2 + \sum_{j=1}^{m}\left(\sum_{i=1}^{n_j}(y_{i,j} - \theta)^2\right)\right]\right) \text{ or}$$

$$\{1/\sigma^2 | \boldsymbol{\theta}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\} \sim \text{gamma}(\frac{1}{2}[\nu_0 + \sum_{j=1}^{m} n_j], \frac{1}{2}[\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2]).$$

It's worth briefly discussing what these values represent. The full conditionals for μ and τ^2 look like standard normal posteriors. Similarly, the full conditional for θ_j looks like a normal posterior dependent only on the specific subgroup $\boldsymbol{y}_j$ and the common variance σ^2. Lastly (and most interestingly), notice that the posterior for σ^2 looks like a standard inverse gamma posterior which depends on $\sum\sum(y_{i,j} - \theta)^2$ which is the pooled variance across all groups (see also $\sum n_j$).

8.4 Example: Math scores in U.S. public schools

8.4.1 Prior distributions and posterior approximation

In this case, we need to specify the following priors:

$$\sigma^2 \sim \text{inverse-gamma}(\nu_0/2, \sigma_0^2\nu_0/2)$$

If we know the math exam was designed to give a nationwide variance of 100, we can set the within-school variance to 100. This is probably an overestimate since the within-school variance should be less than the nationwide estimate. Regardless, we set $\sigma_0^2 = 100, \nu_0 = 1$ to weakly concentrate the prior around 100.

$$\tau^2 \sim \text{inverse-gamma}(\eta_0/2, \tau_0^2\eta_0/2)$$

Similarly, we set $\tau_0^2 = 100, \eta_0 = 1$.

$$\mu^2 \sim \mathcal{N}(\mu_0, \gamma_0^2)$$

Since the mean over all schools should be 50, we set $\mu_0 = 50, \gamma_0^2 = 25$, so that 95% of the probability of our prior is in $(40, 60)$.

Gibbs sampling

Now we sample parameters $\{\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \mu^{(s)}, \tau^{2(s)}, \sigma^{2(s)}\}$, there's a key point about Gibbs sampling that must be emphasized: the order in which we sample the new parameters doesn't matter, but each parameter must be updated according to the most current values of the other parameters: If we have sampled $\mu^{(s+1)}$, the sample of $\tau^{(s+1)}$ must be dependent on $\mu^{(s+1)}$, *NOT* $\mu^{(s)}$. This ensures the Markov chain property.

Given a current state of the unknowns $\{\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \mu^{(s)}, \tau^{2(s)}, \sigma^{2(s)}\}$, a new state is generated as follows:

1. Sample $\mu^{(s+1)} \sim p(\mu|\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \tau^{2(s)})$
2. Sample $\tau^{2(s+1)} \sim p(\tau^2|\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \mu^{(s+1)})$
3. Sample $\sigma^{2(s+1)} \sim p(\sigma^2|\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$
4. for each $j \in \{1, \ldots, m\}$, sample $\theta_j^{(s+1)} \sim p(\theta_j \mid \mu^{(s+1)}, \tau^{2(s+1)}, \sigma^{2(s+1)}, \boldsymbol{y}_j)$

MCMC diagnostics

Before we make inference using these MCMC samples, the first thing we need to do is to see if there are any indications that the chain is not stationary, i.e. if the simulated parameter values are moving in a consistent direction.

Methods:

- Traceplots, or plots of the parameter values versus iteration number (be difficult to read when the number of samples is large)
- Standard practice: plot only a subsequence of MCMC samples, such as every 100th sample
- Produce boxplots of sequential groups of samples (If stationarity has been achieved, then the distribution of samples in any one boxplot should be the same as that in any other)

8.4.2 Posterior summaries and shrinkage

Notice from the full conditional of θ_j above that the expected value of θ_j is a weighted average of $\bar{y}_j$ and μ:

$$\begin{aligned}\mathbb{E}(\theta_j \mid \boldsymbol{y}_j, \mu, \tau^2, \sigma^2) &= \frac{\bar{y}_j n_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2} \\ &= \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu\end{aligned}$$

which is specifically weighted by the sample size n_j. Since we assume that there is some common mean μ, our estimate of θ_j gets pulled slightly towards that common parameter μ- less so for high n_j. In other words, θ_j is pulled a bit from $\bar{y}_j$ towards μ by an amount depending on n_j. This demonstrates the phenomenon of shrinkage, where information is shared across groups in this hierarchical model. Again, for high n_j, however, the effect of this shrinkage is negligible: Groups with low sample sizes get shrunk the most, whereas groups with large sample sizes hardly get shrunk at all. This makes sense: The larger the sample size for a group, the more information we have for that group and the less information we need to "borrow" from the rest of the population.

8.5 Hierarchical modeling of means and variances

The previous model assumed common variance within groups σ^2. This is actually fairly common, perhaps less because of empirical justification for assuming common within-group variance than lack of interest in the variance of the groups, or the mean parameters being of greater interest.

But of course, the inaccuracy of this assumption could result in errors in analysis. It could lead to improper pooling of information, or to the shrinkage of group-specific parameters by inappropriate amounts. It's fairly straightforward to simply add another hierarchical layer for the variance too, and jointly estimate the group-specific means and variances, as well as the common mean and variance parameters.

To implement this, let θ_j depend on $\boldsymbol{y}_j$ and (new!) a group j-specific σ_j^2. Thus, the likelihood (sample function):

$$Y_{1,j}, \ldots, Y_{n_j,j} \sim \text{i.i.d. normal}(\theta_j, \sigma_j^2),$$

and the full conditional distribution is

$$\theta_j \mid \boldsymbol{y}_j, \sigma_j^2 \sim \mathcal{N}\left(\frac{n_j\bar{y}_j/\sigma_j^2 + 1/\tau^2}{n_j/\sigma_j^2 + 1/\tau^2}, \left[n_j/\sigma_j^2 + 1/\tau^2\right]^{-1}\right)$$

Similarly, above we had a rather special case of the full conditional of σ^2. Since there was one common σ^2, the posterior was based on a combination of the prior precision and the pooled sample variance. Now let's assume that we have a separate σ_j^2 for each group:

$$\sigma_1^2, \ldots, \sigma_m^2 \sim \text{i.i.d. Gamma}(\nu_0/2, \sigma_0^2\nu_0/2). \tag{8.2}$$

Include individual σ_j^2, re-derive the full conditional for σ_j^2 results in full conditional distributions that look just like the one-parameter case for variance (and the corresponding θ_j conditionals for the means):

$$\{\sigma_j^2|y_{1,j},\ldots,y_{n_j,j},\theta_j\} = \{\sigma_j^2 \mid \boldsymbol{y}_j,\theta_j\} \sim \text{inverse-gamma}\left([\nu_0+n_j]/2, \left[\nu_0\sigma_0^2+\sum_{i=1}^{n_j}(y_{i,j}-\theta_j)^2\right]/2\right)$$

Discuss: If ν_0 and σ_0^2 are fixed in advance at some particular values: we may obtain $p(\sigma_m^2 \mid \sigma_1^2,\ldots,\sigma_{m-1}^2) = p(\sigma_m^2)$, so the information we may have about $\sigma_m^2 \mid \sigma_1^2,\ldots,\sigma_{m-1}^2$ is not used to help us estimate σ_m^2. This seems inefficient. Thus, we can treat ν_0 and σ_0^2 as parameters to be estimated, in which case (Eq.8.2) is properly thought of as a sampling model for across-group heterogeneity in population variances, and not as a prior distribution. Putting this together with our model for heterogeneity in population means gives a hierarchical model for both means and variances, which is depicted graphically in Figure2.

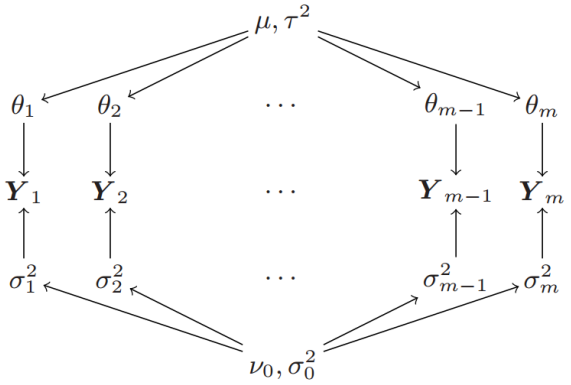


Figure 2: A graphical representation of the hierarchical normal model with heterogeneous means and variances.

The unknown parameters to be estimated include: $\{(\theta_1, \sigma_1^2), \ldots, (\theta_m, \sigma_m^2)\}$ representing the within-group sampling distributions, $\{\mu, \tau^2\}$ representing across-group heterogeneity in means and $\{\nu_0, \sigma_0^2\}$ representing across-group heterogeneity in variances.

As before, the joint posterior distribution for all of these parameters can be approximated by iteratively sampling each parameter from its full conditional distribution given the others.

- full conditional distributions for μ and τ^2 are unchanged from the previous section
- full conditional distributions of θ_j and σ_j^2 are given above
- specify the prior distributions for ν_0 and σ_0^2, and obtain full conditional distributions of ν_0 and σ_0^2
  - A conjugate class of prior densities for σ_0^2 are the gamma densities, i.e., $\sigma_0^2 \sim \text{gamma}(a, b)$. Thus, the posterior is:

$$\sigma_0^2 \mid \sigma_1^2, \ldots, \sigma_m^2, \nu_0 \sim \text{gamma}\left(a + \frac{1}{2}m\nu_0, b + \frac{1}{2}\sum_{j=1}^{m}(1/\sigma_j^2)\right)$$

  - A simple conjugate prior for ν_0 does not exist, but if we restrict ν_0 to be a whole number, then it is easy to sample from its full conditional distribution:
    For example, if we let the prior on ν_0 be the geometric distribution on $\{1, 2, \ldots\}$ so that $p(\nu_0) \propto e^{-\alpha\nu_0}$, then the full conditional distribution of ν_0 is proportional to

$$\begin{aligned} p(\nu_0 \mid \sigma_0^2, \ldots, \sigma_m^2) &\propto p(\nu_0) \times p(\sigma_1^2, \ldots, \sigma_m^2 \mid \nu_0, \sigma_0^2) \\ &\propto \left(\frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\right)^m \left(\prod_{j=1}^{m}\frac{1}{\sigma_j^2}\right)^{\nu_0/2-1} \times \exp\left(-\nu_0\left(\alpha + \frac{1}{2}\sigma_0^2\sum_{j=1}^{m}\frac{1}{\sigma_j^2}\right)\right) \end{aligned}$$

8.6 Discussion and further references

Lindley and Smith (1972) laid the foundation for Bayesian hierarchical modeling, although the idea of shrinking the estimates of the individual group means towards an across-group mean goes back at least to Kelley (1927) in the context of educational testing. In the statistical literature, the benefits of this type of estimation are referred to as the "Stein effect" (Stein, 1956, 1981). Estimators of this type generally take the form $\hat{\theta}_j = w_j \bar{y}_j + (1 - w_j)\bar{y}$, where $\bar{y}$ is an average over all groups and the w_j's depend on n_j, σ^2 and τ^2. So-called empirical Bayes procedures obtain estimates of σ^2 and τ^2 from the data, then plug these values into the formula for $\hat{\theta}_j$ (Efron and Morris, 1973; Casella, 1985). Such procedures often yield estimates of the θ_j's that are nearly equivalent to those from Bayesian procedures, but ignore uncertainty in the values of σ^2 and τ^2. For a detailed treatment of empirical Bayes methods, see Carlin and Louis (1996).

Terminology for hierarchical models is inconsistent in the literature. For the simple hierarchical model $y_{i,j} = \theta_j + \epsilon_{i,j}, \theta_j = \mu + \gamma_j$, the θ_j's (or γ_j's) may be referred to as either "fixed effects" or "random effects," usually depending on how they are estimated. The distribution of the θ_j's is unfortunately often referred to as a prior distribution, which mischaracterizes Bayesian inference and renders the distinction between prior information and population distribution somewhat meaningless.

9 Linear regression

Linear regression modeling is an extremely powerful data analysis tool, useful for a variety of inferential tasks such as prediction, parameter estimation and data description. One difficult aspect of regression modeling is deciding which explanatory variables to include in a model. This variable selection problem has a natural Bayesian solution: Any collection of models having different sets of regressors can be compared via their Bayes factors. When the number of possible regressors is small, this allows us to assign a posterior probability to each regression model. When the number of regressors is large, the space of models can be explored with a Gibbs sampling algorithm.

9.1 The linear regression model

Regression modeling is concerned with describing how the sampling distribution of one random variable Y varies with another variable or set of variables $x = (x_1, \ldots, x_p)$. Specifically, a regression model postulates a form for $p(y|x)$, the conditional distribution of Y given x. Estimation of $p(y|x)$ is made using data $y_1, \ldots, y_n$ that are gathered under a variety of conditions $x_1, \ldots, x_n$.

A linear regression model is a particular type of smoothly changing model for $p(y|\boldsymbol{x})$ that specifies that the conditional expectation $E[Y|\boldsymbol{x}]$ has a form that is linear in a set of parameters:

$$\int yp(y|\boldsymbol{x})dy = \mathrm{E}[Y|\boldsymbol{x}] = \beta_1 x_1 + \cdots + \beta_p x_p = \beta^T \boldsymbol{x}.$$

Thus, actually, $Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i$.

The normal linear regression model specifies that, (1) $E[Y \mid \boldsymbol{x}]$ being linear (2)the sampling variability around the mean is i.i.d. from a normal distribution:

$$Y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i$$
$$\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. normal}(0, \sigma^2)$$

This model provides a complete specification of the joint probability density of observed data $y_1, \ldots, y_n$ conditional upon $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ and values of $\boldsymbol{\beta}$ and σ^2:

$$p(y_1, \ldots, y_n | \boldsymbol{x}_1, \ldots \boldsymbol{x}_n, \beta, \sigma^2) = \prod_{i=1}^{n} p(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) \tag{9.1}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2\}. \tag{9.2}$$

Another way to write the normal regression model is that:

$$\{y | \mathbf{X}, \beta, \sigma^2\} \sim \text{ multivariate normal } (\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix and and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{x}_1 \rightarrow \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \rightarrow \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} \mathrm{E}[Y_1 | \boldsymbol{\beta}, x_1] \\ \vdots \\ \mathrm{E}[Y_n | \boldsymbol{\beta}, x_n] \end{pmatrix}.$$

The density Eq.9.2 depends on $\boldsymbol{\beta}$ through the residuals $y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i$. Given the observed data, the term in the exponent is maximized when the sum of squared residuals, $\text{SSR}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2$ is minimized. Thus, rewrite:

$$\text{SSR}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T x_i)^2 = (y - \mathbf{X}\boldsymbol{\beta})^T(y - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

$$\begin{aligned}
\frac{d}{d\boldsymbol{\beta}}\text{SSR}(\boldsymbol{\beta}) &= \frac{d}{d\beta}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) \\
&= -2\mathbf{X}^T y + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \text{therefore} \\
\frac{d}{d\boldsymbol{\beta}}\text{SSR}(\boldsymbol{\beta}) = 0 &\Leftrightarrow -2\mathbf{X}^T y + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0 \\
&\Leftrightarrow \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T y \\
&\Leftrightarrow \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.
\end{aligned}$$

$\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ is called the "ordinary least squares" (OLS) estimate of $\boldsymbol{\beta}$, as it provides the value of $\boldsymbol{\beta}$ that minimizes the sum of squared residuals.

9.2 Bayesian estimation for a regression model

- We begin with a simple semiconjugate prior distribution for $\boldsymbol{\beta}$ and σ^2 to be used when there is information available about the parameters.
- In situations where prior information is unavailable or difficult to quantify, an alternative "default" class of prior distributions is given.

9.2.1 A semiconjugate prior distribution

The likelihood function, or the sampling density of the data (Eq.9.2), is

$$p(\boldsymbol{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \propto \exp\{-\frac{1}{2\sigma^2}\text{SSR}(\boldsymbol{\beta})\}$$
$$= \exp\{-\frac{1}{2\sigma^2}[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}]\}.$$

First, a multivariate normal prior distribution for $\boldsymbol{\beta}$ is conjugate, then:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \Sigma_0)$$

$$p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\sigma^2) \propto p(\boldsymbol{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \times p(\boldsymbol{\beta})$$
$$\propto \exp\{-\frac{1}{2}(-2\boldsymbol{\beta}^T\mathrm{X}^T\boldsymbol{y}/\sigma^2 + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/\sigma^2) - \frac{1}{2}(-2\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta})\}$$
$$= \exp\{\boldsymbol{\beta}^T(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\boldsymbol{y}/\sigma^2) - \frac{1}{2}\boldsymbol{\beta}^T(\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)\boldsymbol{\beta}\}.$$

Referring Chapter 7, it is easy to recognize that this as being proportional to a multivariate normal density, with

$$\text{Var}[\boldsymbol{\beta}|y,\mathbf{X},\sigma^2] = (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1} \tag{9.3}$$
$$\text{E}[\boldsymbol{\beta}|y,\mathbf{X},\sigma^2] = (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\boldsymbol{y}/\sigma^2) \tag{9.4}$$

- If Σ_0^{-1} small: $E[\boldsymbol{\beta}|y,\mathbf{X},\sigma^2]$ approximately equal to $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$, the least squares estimate.
- If σ^2 large (measurement precision is small): $E[\boldsymbol{\beta}|y,\mathbf{X},\sigma^2]$ approximately equal to $\boldsymbol{\beta}_0$, the prior expectation.

Second, a inverse-gamma distribution for σ^2 is conjugate. $\gamma = 1/\sigma^2$, the measurement precision, then:

$$\gamma \sim \text{gamma}(\nu_0/2, \nu\sigma^2/2)$$

$$\begin{aligned} p(\gamma|\boldsymbol{y}, \mathbf{X}, \boldsymbol{\beta}) &\propto p(\gamma)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \gamma) \\ &\propto \left[\gamma^{\nu_0/2-1} \exp(-\gamma \times \nu_0\sigma_0^2/2)\right] \times \left[\gamma^{n/2} \exp(-\gamma \times \text{SSR}(\boldsymbol{\beta})/2)\right] \\ &= \gamma^{(\nu_0+n)/2-1} \exp(-\gamma[\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})]/2), \\ \{\sigma^2|y, \mathbf{X}, \boldsymbol{\beta}\} &\sim \text{inverse-gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + SSR(\boldsymbol{\beta})]/2). \end{aligned}$$

Then, constructing a Gibbs sampler to approximate the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \mathbf{X})$ is then straightforward: Given current values $\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}\}$, new values can be generated by

1. updating $\boldsymbol{\beta}$
   a) compute $\mathbf{V} = \text{Var}[\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \sigma^{2(s)}]$ and $\mathbf{m} = \text{E}[\boldsymbol{\beta}|y, \mathbf{X}, \sigma^{2(s)}]$
   b) sample $\beta^{(s+1)} \sim \text{multivariate normal}(\mathbf{m}, \mathbf{V})$
2. updating σ^2
   a) compute $\text{SSR}(\boldsymbol{\beta}^{(s+1)})$
   b) sample $\sigma^{2(s+1)} \sim \text{inverse} - \text{gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta}^{(s+1)})]/2)$

9.2.2 Default and weakly informative prior distributions

A Bayesian analysis of a regression model requires specification of the prior parameters $(\boldsymbol{\beta}_0, \Sigma_0)$ and (ν_0, σ_0^2). Finding values of these parameters that represent actual prior information can be difficult.

The task of constructing an informative prior distribution only gets harder as the number of regressors increases, as the number of prior correlation parameters is $\binom{p}{2}$, which increases quadratically in p.

One idea is that, if the prior distribution is not going to represent real prior information about the parameters, then it should be as minimally informative as possible. The resulting posterior distribution would then represent the posterior information of someone who began with little knowledge of the population being studied. Such an analysis would give a "more objective" result than using an informative prior distribution, especially one that did not actually represent real prior information.

1. One type of weakly informative prior is the *unit information prior*. A unit information prior is one that contains the same amount of information as that would be contained in only a single observation.
   - For example, the precision of $\hat{\boldsymbol{\beta}}_{\text{ols}}$ is its inverse variance, or $(\mathbf{X}^T\mathbf{X})/(\sigma^2)$. Since this can be viewed as the amount of information in n observations, the amount of information in one observation should be "one nth" as much, i.e.$(\mathbf{X}^T\mathbf{X})/(n\sigma^2)$. The unit information prior thus sets $\Sigma_0^{-1} = (\mathbf{X}^T\mathbf{X})/(n\sigma^2)$.
   - If setting $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{\text{ols}}$ thus centering the prior distribution of $\boldsymbol{\beta}$ around the OLS estimate: Such a distribution cannot be strictly considered a real prior distribution, as it requires knowledge of $\boldsymbol{y}$ to be constructed. However, it only uses a small amount of the information in $\boldsymbol{y}$, and can be loosely thought of as the prior distribution of a person with unbiased but weak prior information.
   - In a similar way, the prior distribution of σ^2 can be weakly centered around $\hat{\sigma}^2$ ols by taking $\nu_0 = 1$ and $\sigma_0^2 = \hat{\sigma}^2_{\text{ols}}$.

2. **g-prior**

Idea: the parameter estimation should be invariant to changes in the scale of the regressors.

Suppose $\mathbf{X}$ is a given set of regressors and $\tilde{\mathbf{X}} = \mathbf{XH}$ for some $p \times p$ matrix $\mathbf{H}$. If we obtain the posterior distribution of $\boldsymbol{\beta}$ from $\boldsymbol{y}$ and $\mathbf{X}$, and the posterior distribution of $\tilde{\beta}$ from $\boldsymbol{y}$ and $\tilde{\boldsymbol{X}}$, then, according to this principle of invariance, the posterior distributions of $\boldsymbol{\beta}$ and $\mathbf{H}\tilde{\boldsymbol{\beta}}$ should be the same.

This condition will be met if $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\Sigma_0 = k(\mathbf{X}^T\mathbf{X})^{-1}$ for any positive value k. A popular specification of k is to relate it to the error variance σ_2 , so that $k = g\sigma_2$ for some positive value g.

These choices of prior parameters result in a version of the so-called "g-prior".

Under this invariant g-prior the conditional distribution of $\boldsymbol{\beta}$ given $(\boldsymbol{y}, \mathbf{X}, \sigma^2)$ is still multivariate normal, but Eqs.9.3 and 9.4 reduce to the following simpler forms:

$$\mathrm{Var}[\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \sigma^2] = [\mathbf{X}^T\mathbf{X}/(g\sigma^2) + \mathbf{X}^T\mathbf{X}/\sigma^2]^{-1} = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{9.5}$$

$$\mathrm{E}[\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \sigma^2] = [\mathbf{X}^T\mathbf{X}/(g\sigma^2) + \mathbf{X}^T\mathbf{X}/\sigma^2]^{-1}\mathbf{X}^T y/\sigma^2 = \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}. \tag{9.6}$$

Parameter estimation under the g-prior is simplified as well: It turns out that, under this prior distribution, $p(\sigma^2 \mid \boldsymbol{y}, \mathbf{X})$ is an inverse-gamma distribution, which means that we can directly sample $(\sigma^2, \boldsymbol{\beta})$ from their posterior distribution by first sampling from $p(\sigma^2 \mid \boldsymbol{y}, \mathbf{X})$ and then from $p(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}, \mathbf{X})$.

Derivation of $p(\sigma^2|\boldsymbol{y}, \mathbf{X})$

The marginal posterior density of σ^2 is proportional to $p(\sigma^2) \times p(\boldsymbol{y}|\mathbf{X}, \sigma^2)$. Using the rules of marginal probability, the latter term in this product can be expressed as the following integral:

$$p(\boldsymbol{y}|\mathbf{X},\sigma^2) = \int p(\boldsymbol{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|\mathbf{X},\sigma^2)\, d\boldsymbol{\beta}.$$

Writing out the two densities inside the integral, we have

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|\mathbf{X},\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]\times \\ |2\pi g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}|^{-1}\exp[-\frac{1}{2g\sigma^2}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}]. \end{aligned} \tag{9.7}$$

Combining the terms in the exponents gives

$$\begin{aligned} &-\frac{1}{2\sigma^2}\left[(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/g\right] \\ &= -\frac{1}{2\sigma^2}\left[\boldsymbol{y}^T\boldsymbol{y}-2y^T\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}(1+1/g)\right] \\ &= -\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})^T\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})+\frac{1}{2}\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}, \end{aligned}$$

$$\text{where } \mathbf{V} = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \quad \text{and} \quad \mathbf{m} = \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}.$$

This means that we can write Eq.9.7 as

$$\begin{aligned} &\left[(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y})\right]\times\left[(1+g)^{-p/2}\exp(\frac{1}{2}\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m})\right]\times \\ &\left[|2\pi\mathbf{V}|^{-1/2}\exp[-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})^T\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})]\right]. \end{aligned}$$

The third term in the product is the only term that depends on $\boldsymbol{\beta}$. This term is exactly the multivariate normal density with mean $\mathbf{m}$ and variance $\mathbf{V}$, which as a probability density must integrate to 1. This means that if we integrate the whole thing with respect to $\boldsymbol{\beta}$ we are left with only the first two terms:

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X},\sigma^2) &= \int p(\boldsymbol{y}|\boldsymbol{\beta},\mathbf{X})p(\boldsymbol{\beta}|\mathbf{X},\sigma^2)d\boldsymbol{\beta} \\ &= \left[(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y})\right] \times \left[(1+g)^{-p/2}\exp(\frac{1}{2}\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m})\right] \times 1, \end{aligned} \tag{9.8}$$

which, after combining the terms in the exponents, is

$$p(\boldsymbol{y}|\mathbf{X},\sigma^2) = (2\pi)^{-n/2}(1+g)^{-p/2}(\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\mathrm{SSR}_g),$$

where SSR_g is defined as

$$\mathrm{SSR}_g = \boldsymbol{y}^T\boldsymbol{y} - \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} = \boldsymbol{y}^T(\mathbf{I} - \frac{g}{g+1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\boldsymbol{y}.$$

As $g \to \infty$, SSR_g decreases to $\mathrm{SSR}_{\mathrm{ols}} = \sum(y_i - \hat{\boldsymbol{\beta}}_{\mathrm{ols}}\boldsymbol{x}_i)^2$. The effect of g is that it shrinks down the magnitude of the regression coefficients and can prevent overfitting of the data.

The last step in identifying $p(\sigma^2 \mid \boldsymbol{y}, \mathbf{X})$ is to multiply $p(\boldsymbol{y} \mid \mathbf{X}, \sigma^2)$ by the prior distribution. Letti $\gamma = 1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$, we have

$$
\begin{aligned}
p(\gamma|\boldsymbol{y},\mathbf{X}) &\propto p(\gamma)p(\boldsymbol{y}|\mathbf{X},\gamma) \\
&\propto \left[\gamma^{\nu_0/2-1}\exp(-\gamma\times\nu_0\sigma_0^2/2)\right]\times\left[\gamma^{n/2}\exp(-\gamma\times \mathrm{SSR}_g/2)\right] \\
&= \gamma^{(\nu_0+n)/2-1}\exp[-\gamma\times(\nu_0\sigma_0^2+\mathrm{SSR}_g)/2] \\
&\propto \mathrm{dgamma}(\gamma,[\nu_0+n]/2,[\nu_0\sigma_0^2+\mathrm{SSR}_g]/2),
\end{aligned}
$$

and $\{\sigma^2|\boldsymbol{y},\mathbf{X}\}\sim$ inverse-gamma$([\nu_0+n]/2,[\nu_0\sigma_0^2+\mathrm{SSR}_g]/2)$.

Under g-prior distribution, $p(\sigma^2 \mid \boldsymbol{y},\mathbf{X})$ and $p(\boldsymbol{\beta} \mid \boldsymbol{y},\mathbf{X},\sigma^2)$ are inverse-gamma and multivariate normal distributions respectively. Since we can sample from both of these distributions, samples from the joint posterior distribution $p(\sigma^2,\boldsymbol{\beta} \mid \boldsymbol{y},\mathbf{X})$ can be made with Monte Carlo approximation, and Gibbs sampling is unnecessary.

A sample value of $(\sigma^2,\boldsymbol{\beta})$ from $p(\sigma^2,\boldsymbol{\beta} \mid \boldsymbol{y},\mathbf{X})$ can be made as follows:

1. sample $1/\sigma^2\sim$ gamma$([\nu_0+n]/2,[\nu_0\sigma_0^2+\mathrm{SSR}_g]/2)$
2. sample $\boldsymbol{\beta}\sim$ multivariate normal$(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}},\frac{g}{g+1}\sigma^2[\mathbf{X}^T\mathbf{X}]^{-1})$

9.3 Model selection

We should include in our regression model only those variables for which there is substantial evidence of an association with y.

Perhaps our predictions could be improved by *removing from the regression model those variables that show little evidence of being nonzero*. By doing so, we hope to remove from the predictive model any regressors that have spurious associations to Y (i.e. those associations specific only to the training data), leaving only those regressors that would have associations for any group of subjects (i.e. both the training and test data).

One standard way to assess the evidence that the true value of a regression coefficient β_j is not zero is with a t-statistic, which is obtained by dividing the OLS estimate $\hat{\beta}_j$ by its standard error, so $t_j = \hat{\beta}_j / [\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}_{j,j}]^{1/2}$.

A procedure, in which a potentially large set of regressors is reduced to a smaller set, are called *model selection procedures*. Consider the following procedure:

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$
   - find the regressor j_{min} having the smallest value of $|t_j|$ and remove column j_{min} from $\mathbf{X}$.
   - return to Step 1
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

A standard choice for t_{cutoff} is an upper quantile of a t or standard normal distribution.

9.3.1 Bayesian model comparison

The Bayesian solution to the model selection problem is conceptually straightforward:

If we believe that many of the regression coefficients are potentially equal to zero, then we simply come up with a prior distribution that reflects this possibility. This can be accomplished by specifying that each regression coefficient has some non-zero probability of being exactly zero.

A convenient way to represent this is to write the regression coefficient for variable j as $\beta_j = z_j \times b_j$, where $z_j \in \{0, 1\}$ and b_j is some real number (The z_j's indicate which regression coefficients are non-zero). With this parameterization, our regression equation becomes:

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

Each value of $\boldsymbol{z} = (z_1, \ldots, z_p)$ corresponds to a different model. (Eg: $\boldsymbol{z} = (1, 0, 1, 0)$ and $\boldsymbol{z} = (1, 1, 1, 0)$ correspond to different models.) With this parameterization, choosing which variables to include in a regression model is equivalent to choosing which z_j's are 0 and which are 1.

Bayesian model selection proceeds by obtaining a posterior distribution for $\boldsymbol{z}$. Of course, doing so requires a joint prior distribution on $\{\boldsymbol{z}, \boldsymbol{\beta}, \sigma^2\}$.

It turns out that a version of the g-prior allows us to evaluate $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z})$ for each possible model $\boldsymbol{z}$.

Given a prior distribution $p(\boldsymbol{z})$ over models, this allows us to compute a posterior probability for each regression model:

$$p(\boldsymbol{z}|\boldsymbol{y}, \mathbf{X}) = \frac{p(\boldsymbol{z})p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z})}{\sum_{\tilde{\boldsymbol{z}}} p(\tilde{\boldsymbol{z}})p(\boldsymbol{y}|\mathbf{X}, \tilde{\boldsymbol{z}})}.$$

Alternatively, we can compare the evidence for any two models with the posterior odds:

$$\begin{aligned} \text{odds}(\boldsymbol{z}_a, \boldsymbol{z}_b|\boldsymbol{y}, \mathbf{X}) &= \frac{p(\boldsymbol{z}_a|\boldsymbol{y}, \mathbf{X})}{p(\boldsymbol{z}_b|\boldsymbol{y}, \mathbf{X})} = \frac{p(\boldsymbol{z}_a)}{p(\boldsymbol{z}_b)} \times \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_a)}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_b)} \\ \text{posterior odds} &= \text{prior odds} \times \text{“Bayes factor”} \end{aligned} \tag{9.9}$$

The Bayes factor can be interpreted as how much the data favor model $\boldsymbol{z}_a$ over model $\boldsymbol{z}_b$.

Computing the marginal probability

In order to obtain a posterior distribution over models, we will have to compute $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z})$ for each model $\boldsymbol{z}$ under consideration. The marginal probability is obtained from the integral

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}) &= \int\int p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2|\mathbf{X}, \boldsymbol{z}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int\int p(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{X}) p(\boldsymbol{\beta}|\mathbf{X}, \boldsymbol{z}, \sigma^2) p(\sigma^2) d\boldsymbol{\beta} d\sigma^2. \end{aligned} \tag{9.10}$$

Using a version of the g-prior distribution for $\boldsymbol{\beta}$, we will be able to compute this integral without needing much calculus. For any given $\boldsymbol{z}$ with p_z non-zero entries, let $\mathbf{X}_z$ be the $n \times p_z$ matrix corresponding to the variables j for which $z_j = 1$, and similarly let $\boldsymbol{\beta}_z$ be the $p_z \times 1$ vector consisting of the entries of $\boldsymbol{\beta}$ for which $z_j = 1$.

Our modified g-prior distribution for $\boldsymbol{\beta}$ is that $\beta_j = 0$ for j's such that $z_j = 0$, and that

$$\{\boldsymbol{\beta}_z|\mathbf{X}_z, \sigma^2\} \sim \text{multivariate normal}(\mathbf{0}, g\sigma^2[\mathbf{X}_z^T\mathbf{X}_z]^{-1}).$$

Intergrate Eq.9.10 w.r.t $\boldsymbol{\beta}$ first, we have:

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}) &= \int \left(\int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \sigma^2, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{X}, \boldsymbol{z}, \sigma^2) d\boldsymbol{\beta} \right) p(\sigma^2) d\boldsymbol{\sigma}^2 \\ &= \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \sigma^2) p(\sigma^2) d\sigma^2. \end{aligned}$$

We can obtain the calculation about $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \sigma^2)$ in Eq.9.8. Then, define $\gamma = 1/\sigma^2$, let $p(\gamma)$ be the gamma density with parameters $(\nu_0/2, \nu_0\sigma_0^2/2)$, we can show that conditional density of $(\boldsymbol{y}, \gamma)$ given $(\mathbf{X}, \boldsymbol{z})$ is

$$\begin{aligned} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}) = p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}, \gamma) \times p(\gamma) = (2\pi)^{-n/2}(1+g)^{-p_z/2} \times \left[\gamma^{n/2} e^{-\gamma \mathrm{SSR}_g^z/2} \right] \times \\ (\nu_0\sigma_0^2/2)^{\nu_0/2} \Gamma(\nu_0/2)^{-1} \left[\gamma^{\nu_0/2-1} e^{-\gamma\nu_0\sigma_0^2/2} \right], \end{aligned} \tag{9.11}$$

where SSR_g^z is as before, based on the regressor matrix $\mathbf{X}_z$:

$$\mathrm{SSR}_g^z = y^T(\mathbf{I} - \frac{g}{g+1}\mathbf{X}_z(\mathbf{X}_z^T\mathbf{X}_z)^{-1}\mathbf{X}_z)\boldsymbol{y}.$$

The part of Eq.9.11 that depends on γ is proportional to a gamma density, but in this case the normalizing constant is the part that we need:

$$\begin{aligned} &\gamma^{(\nu_0+n)/2-1} \exp[-\gamma \times (\nu_0\sigma_0^2 + \mathrm{SSR}_g^z)/2] = \\ &\frac{\Gamma([\nu_0+n]/2)}{([\nu_0\sigma_0^2 + \mathrm{SSR}_g^z]/2)^{(\nu_0+n)/2-1}} \times \mathrm{dgamma}[\gamma, (\nu_0+n)/2, (\nu_0\sigma_0^2 + \mathrm{SSR}_g^z)/2] \end{aligned}$$

Thus we can have:

$$p(y|\mathbf{X}, z) = \pi^{-n/2}\frac{\Gamma([\nu_0+n]/2)}{\Gamma(\nu_0/2)}(1+g)^{-p_z/2}\frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2+\mathrm{SSR}_g^z)^{(\nu_0+n)/2}}.$$

Suppose $g = n$ and use the unit information prior for $p(\sigma^2)$ for each model $\boldsymbol{z}$, so that $\nu_0 = 1$ for all $\boldsymbol{z}$, but σ_0^2 is the estimated residual variance under the least squares estimate for model $\boldsymbol{z}$.

Recall Eq.9.9. In this case, the ratio of the probabilities under any two models $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$ is

$$\frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_a)}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_b)} = (1+n)^{(p_{zb}-p_{za})/2}\left(\frac{s_{z_a}^2}{s_{zb}^2}\right)^{1/2} \times \left(\frac{s_{z_b}^2+\mathrm{SSR}_g^{z_b}}{s_{z_a}^2+\mathrm{SSR}_g^{z_a}}\right)^{(n+1)/2}.$$

Notice that the ratio of the marginal probabilities is essentially a balance between model complexity and goodness of fit: A large value of p_{z_b} compared to p_{z_a} penalizes model $\boldsymbol{z}_b$, although a large value of $\mathrm{SSR}_g^{z_a}$ compared to $\mathrm{SSR}_g^{z_b}$ penalizes model $\boldsymbol{z}_a$.

- $p_{z_b} \uparrow$, more variables, more complexity, large $\frac{p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z}_a)}{p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z}_b)}$, $\boldsymbol{z}_a$✓
- $\mathrm{SSR}_g^{z_a} \uparrow$, less goodness of fit, small $\frac{p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z}_a)}{p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z}_b)}$, $\boldsymbol{z}_b$✓

9.3.2 Gibbs sampling and model averaging

If we allow each of the p regression coefficients to be either zero or non-zero, then there are 2^p different models to consider: impractical for us to compute the marginal probability of each model.

Our data analysis goals become more modest: For example, we may be content with a decent estimate of $\boldsymbol{\beta}$ from which we can make predictions, or a list of relatively high-probability models. These items can be obtained with a Markov chain which searches through the space of models for values of $\boldsymbol{z}$ with high posterior probability. This can be done with a Gibbs sampler in which we iteratively sample each z_j from its full conditional distribution.

Specifically, given a current value $\boldsymbol{z} = (z_1, \ldots, z_p)$, a new value of z_j is generated by sampling from $p(z_j|\boldsymbol{y}, \mathbf{X}, \boldsymbol{z}_{-j})$. The full conditional probability that z_j is 1 can be written as $o_j/(1 + o_j)$, where o_j is the conditional odds that z_j is 1, given by

$$o_j = \frac{\Pr(z_j = 1|\boldsymbol{y}, \mathbf{X}, \boldsymbol{z}_{-j})}{\Pr(z_j = 0|\boldsymbol{y}, \mathbf{X}, \boldsymbol{z}_{-j})} = \frac{\Pr(z_j = 1)}{\Pr(z_j = 0)} \times \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_{-j}, z_j = 1)}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_{-j}, z_j = 0)}.$$

We also want to obtain posterior samples of $\boldsymbol{\beta}$ and σ^2. Using the previous results, these parameters can be sampled directly from their conditional distributions given $\boldsymbol{z}$, $\boldsymbol{y}$ and $\mathbf{X}$:

For each $\boldsymbol{z}$ in our MCMC sample,we can construct the matrix $\mathbf{X}_z$ which consists of only those columns j corresponding to non-zero values of z_j. Using this matrix of regressors, sample σ^2 from $p(\sigma^2 \mid \mathbf{X}, \boldsymbol{y}, \boldsymbol{z})$ (inverse-gamma) and $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta} \mid \mathbf{X}, \boldsymbol{y}, \boldsymbol{z}, \sigma^2)$ (multivariate normal). Our Gibbs sampling scheme therefore looks Something like the following:

$$\begin{array}{l} \boldsymbol{z}^{(s)} \longrightarrow \sigma^{2(s)} \longrightarrow \boldsymbol{\beta}^{(s)} \\ \downarrow \\ \boldsymbol{z}^{(s+1)} \longrightarrow \sigma^{2(s+1)} \longrightarrow \boldsymbol{\beta}^{(s+1)} \end{array}$$

Generating values of $\{\boldsymbol{z}^{(s+1)}, \sigma^{(s+1)}, \boldsymbol{\beta}^{(s+1)}\}$ from $\boldsymbol{z}^{(s)}$ is achieved with the following steps:

1. Set $\boldsymbol{z} = \boldsymbol{z}^{(s)}$;
2. For $j \in \{1, \ldots, p\}$ in random order, replace z_j with a sample from $p(z_j | \boldsymbol{z}_{-j}, \boldsymbol{y}, \mathbf{X})$
3. Set $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}$;
4. Sample $\sigma^{2(s+1)} \sim p(\sigma^2 | \boldsymbol{z}^{(s+1)}, y, \mathbf{X})$
5. Sample $\boldsymbol{\beta}^{(s+1)} \sim p(\boldsymbol{\beta} | \boldsymbol{z}^{(s+1)}, \sigma^{2(s+1)}, \mathbf{y}, \mathbf{X})$

Note that the entries of $\boldsymbol{z}^{(s+1)}$ are not sampled from their full conditional distributions given $\sigma^{2(s)}$ and $\boldsymbol{\beta}^{(s)}$. This is not a problem: The Gibbs sampler for $\boldsymbol{z}$ ensures that the distribution of $\boldsymbol{z}^{(s)}$ converges to the target posterior distribution $p(\boldsymbol{z} | \boldsymbol{y}, \mathbf{X})$. Since $(\sigma^{2(s)}, \boldsymbol{\beta}^{(s)})$ are direct samples from $p(\sigma^2, \boldsymbol{\beta} | \boldsymbol{z}^{(s)}, \boldsymbol{y}, \mathbf{X})$, the distribution of $(\sigma^{2(s)}, \boldsymbol{\beta}^{(s)})$ converges to $p(\sigma^2, \boldsymbol{\beta} | \boldsymbol{y}, \mathbf{X})$.

9.4 Discussion and further references

Many have argued that in most situations none of the regression models under consideration are actually true. In this situation, Bayesian model selection can still be meaningful in a decision-theoretic sense, where the task is to select the model with the best predictive performance. In this case, model selection proceeds using a modified Bayes factor that is similar to a cross-validation criterion.

10 Nonconjugate priors and Metropolis-Hastings algorithms

We present the Metropolis-Hastings algorithm as a generic method of approximating the posterior distribution corresponding to any combination of prior distribution and sampling model.

Two examples: (1) Poisson regression, a type of generalized linear model. (2) A longitudinal regression model in which the observations are correlated over time

10.1 Generalized linear models

We can use an example to understand the Generalized Linear models:

For Song sparrow, denote Y, the number of offspring conditional, and x, their ages. A simple probability model would be a Poisson model:

$$\{Y \mid x\} \sim \text{Poisson}(\theta_x)$$

Then we can estimate θ_x for each age group. But the number of birds of each age is small. To add stability to the estimation we will assume that the mean number of offspring is a smooth function of age. We will want to allow this function to be quadratic. One possibility would be to express θ_x as $\theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$. To make sure $\theta_x > 0$, model the log-mean of Y in terms of this regression, so that

$$\log \mathrm{E}[Y|x] = \log \theta_x = \beta_1 + \beta_2 x + \beta_3 x^2$$

and $\mathrm{E}[Y|x] = \log \theta_x = \exp(\beta_1 + \beta_2 x + \beta_3 x^2)$.

Noted that we have different ages (different x). Then $\log E[Y_i|x_i] = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$, where x_i is the age of the sparrow i. Denote $\boldsymbol{x}_i = (1, x_i, x_i^2)$ so that $\log E[Y_i|x_i] = \boldsymbol{\beta}^T \boldsymbol{x}_i$.

Thus, the resulting model is:

$$\{Y|\boldsymbol{x}\} \sim \text{Poisson}(\exp[\boldsymbol{\beta}^T \boldsymbol{x}])$$

which is called *Poisson regression model*. The term $\boldsymbol{\beta}^T$ is *linear predictor*, which linked to $E[Y|\boldsymbol{x}]$ via the log function so we say this model has a *log link*.

- The Poisson regression model is a type of generalized linear model, a model which relates a function of the expectation to a linear predictor of the form $\boldsymbol{\beta}^T \boldsymbol{x}$.
- Another common generalized linear model is the logistic regression model for binary data. Writing $Pr(Y = 1|\boldsymbol{x}) = E[Y|\boldsymbol{x}] = \theta_x$, the logistic regression model parameterizes θ_x as

$$\theta_x = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x})}, \text{ so that } \boldsymbol{\beta}^T \boldsymbol{x} = \log \frac{\theta_x}{1 - \theta_x}$$

  The function $\log \theta_x/(1 - \theta_x)$ relating the mean to the linear predictor is called the *logit function*, so the logistic regression model could be described as *a binary regression model with a logit link*.

Although independent Monte Carlo sampling from the posterior is not available for this Poisson regression model, the next section will show how to construct a Markov chain that can approximate $p(\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y})$ for any prior distribution $p(\boldsymbol{\beta})$.

10.2 The Metropolis algorithm

- (A generic situation:) we have a sampling model $Y \sim p(y|\theta)$ and a prior distribution $p(\theta)$. Posterior $p(\theta|y) = p(\theta)p(y|\theta)/\int p(\theta')p(y|\theta')d\theta'$ is often hard to calculate due to the integral. If we were able to sample from $p(\theta|y)$, then we generate $\theta^{(1)}, \ldots, \theta^{(S)} \sim$ i.i.d.$p(\theta|y)$, and can obtain Monte Carlo approximations to posterior quantities, such as $\mathrm{E}[g(\theta)|y] \approx \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$.

What if we cannot sample directly from $p(\theta|y)$? Actually, in terms of approximating the posterior distribution, the critical thing is not that we have i.i.d. samples from $p(\theta|y)$ but rather that we are able to construct a large collection of θ-values, $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$, whose empirical distribution approximates $p(\theta|y)$.

Suppose we have a working collection $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ to which we would like to add a new value $\theta^{(s+1)}$. We consider adding a value θ^* which is nearby $\theta^{(s)}$. If $p(\theta^*|y) > p(\theta^{(s)}|y)$, we should include θ^* as well. If $p(\theta^*|y) < p(\theta^{(s)}|y)$? This comparison can be made even if we cannot compute $p(\theta|y)$

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)} \frac{p(y)}{p(y|\theta^{(s)})p(\theta^{(s)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}. \tag{10.1}$$

- If $r > 1$:

  Intuition: Since $\theta^{(s)}$ is already in our set, we should include θ^* as it has a higher probability than $\theta^{(s)}$.

  Procedure: Accept θ^* into our set, i.e. set $\theta^{(s+1)} = \theta^*$.
- If $r < 1$:

  Intuition: The relative frequency of θ in our set equal to θ^* compared to those equal to $\theta^{(s)}$ should be $p(\theta^*|y)/p(\theta^{(s)}|y) = r$. For each of $\theta^{(s)}$, we should have only a "fraction"of an instance of a θ^* value.

  Procedure: Set $\theta^{(s+1)}$ equal to either θ^* or $\theta^{(s)}$, with probability r and $1 - r$ respectively.

10.2.1 Metropolis algorithm

The Metropolis algorithm proceeds by sampling a proposal value θ^* nearby the current value $\theta^{(s)}$ using a symetric proposal distribution $J(\theta^*|\theta^{(s)})$.

- Symmetric: $J(\theta_b|\theta_a) = J(\theta_a|\theta_b)$. Eg: $J(\theta^*|\theta^{(s)}) = \text{uniform}(\theta^{(s)}-\delta, \theta^{(s)}+\delta)$; $J(\theta^*|\theta^{(s)}) = \text{normal}(\theta^{(s)}, \delta^2)$
- Noted that the value of the parameter δ is chosen to make the approximation algorithm run efficiently.
- Give $\theta^{(s)}$, Metropolis algorithm generates a value $\theta^{(s+1)}$ as follows:
  1. Sample $\theta^* \sim J(\theta|\theta^{(s)})$
  2. Compute the acceptance ratio: $r = \dfrac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \dfrac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}$.
  3. Let

$$\theta^{(s+1)} = \begin{cases} \theta^* \text{ with probability } \min(r, 1) \\ \theta^{(s)} \text{ with probability } 1 - \min(r, 1). \end{cases}$$

Accomplish Step 3: Sample $u \sim$uniform$(0, 1)$, set $\theta^{(s+1)} = \theta^*$ if $u < r$ and $\theta^{(s+1)} = \theta^{(s)}$ otherwise.

10.2.2 Output of the Metropolis algorithm

MA generates a dependent sequence $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ of θ-values. Since $\theta^{(s+1)}$ depends only on $\theta^{(s)}$, the conditional distribution of $\theta^{(s+1)}$ given $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ also depends only on $\theta^{(s)}$ and so the sequence $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ is a Markov chain. Under some mild conditions:

- The marginal sampling distribution of $\theta(s)$ is approximately $p(\theta|y)$ for large s.
- For any given numerical value θ_a of θ, $\lim\limits_{S\to\infty} \#\{\theta\text{'s in the sequence } < \theta_a\}/S = p(\theta < \theta_a|y)$.

However, in practice, using either the Metropolis algorithm or the Gibbs sampler, we don't use $s \to \infty$ but follow the standard practice in MCMC approximation:

1. run algorithm until some iteration B for which it looks like the Markov chain has achieved stationarity
2. run the algorithm S more times, generating $\{\theta^{(B+1)}, \ldots, \theta^{(B+S)}\}$
3. discard $\{\theta^{(1)}, \ldots, \theta^{(B)}\}$ and use the empirical distribution of $\{\theta^{(B+1)}, \ldots, \theta^{(B+S)}\}$ to approximate $p(\theta|y)$

The iterations up to and including B: the "burn-in" period. In which the Markov chain moves from its initial value to a region of the parameter space that has high posterior probability.

correlation or δ

The θ-values generated from an MCMC algorithm are statistically dependent. Recall MCMC diagnostics in Chapter 6: the higher the correlation, the longer it will take for the Markov chain to achieve stationarity. Because the amount of information we obtain about $E[\theta|y]$ from S positively correlated samples is less than the information we would obtain from S independent samples. The more correlated our Markov chain is, the less information we get per iteration.

With the Metropolis algorithm the correlation can be adjusted by selecting an optimal value of δ in the proposal distribution.

- It is common practice to first select a proposal distribution by implementing several short runs of the Metropolis algorithm under different δ-values until one is found that gives an acceptance rate roughly between 20 and 50%. Once a reasonable δ is selected, a longer more efficient Markov chain can be run.
- Modified versions of the Metropolis algorithm can be constructed that adaptively change the value of δ at the beginning of the chain in order to automatically find a good proposal distribution.

10.2.3 The Metropolis algorithm for Poisson regression

The prior distribution we used was that the regression coeficients were i.i.d. normal(0,100).

Given a current value $\boldsymbol{\beta}^{(s)}$ and a value $\boldsymbol{\beta}^*$ generated from $J(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(s)})$,the acceptance ratio for the Metropolis algorithm is

$$r = \frac{p(\beta^*|\mathbf{X}, \boldsymbol{y})}{p(\beta^{(s)}|\mathbf{X}, \boldsymbol{y})} = \frac{\prod_{i=1}^n \text{dpois}(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta}^*)}{\prod_{i=1}^n \text{dpois}(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta}^{(s)})} \times \frac{\prod_{j=1}^3 \text{dnorm}(\beta_j^*, 0, 10)}{\prod_{j=1}^3 \text{dnorm}(\beta_j^{(s)}, 0, 10)}.$$

Then what we need is to specify the proposal distribution for θ^*. In many problems, the posterior variance can be an efficient choice of a proposal variance. Although we do not know it before runing MA, a rough approximation is often sufficient.

A convenient choice is a multivariate normal distribution with mean $\boldsymbol{\beta}^{(s)}$. In a normal regression problem, the posterior variance of $\boldsymbol{\beta}$ will be close to $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, where σ^2 is the variance of Y. In our Poisson regression, the model is that the log of Y has expectation equal to $\boldsymbol{\beta}^T\boldsymbol{x}$, so let's try a proposal variance of $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ where $\hat{\sigma}^2$ is the sample variance of $\{\log(y_1 + 1/2), \ldots, \log(y_n + 1/2)\}$ (we use $\log(y + 1/2)$ instead of log$y$ because the latter would be $-\infty$ if $y = 0$).

10.3 Metropolis, Metropolis-Hastings and Gibbs

A Markov chain is a sequentially generated sequence $\{x^{(1)}, x^{(2)}, \ldots\}$ such that the mechanism that generates a $x^{(s+1)}$ can depend on the value of $x^{(s)}$ but not on $\{x^{(s-1)}, x^{(s-2)}, \ldots x^{(1)}\}$.

The Gibbs Sampler and the Metropolis algorithm are both ways of generating Markov chains that approximate a target probability distribution $p_0(x)$ for a potentially vector-valued random variable x. In Bayesian analysis x is typically a parameter or vector of parameters and $p_0(x)$ is a posterior distribution.

These two algorithms are in fact special cases of a more general algorithm, called the Metropolis-Hastings algorithm. Markov chains generated by the Metropolis-Hastings algorithm are able to approximate a target probability distribution.

10.3.1 The Metropolis-Hastings algorithm

We'll first consider a simple example where our target probability distribution is $p_0(u, v)$, a bivariate distribution for two random variables U and V.

- **Gibbs**
  Given $x^{(s)} = (u^{(s)}, v^{(s)})$, a new value of $x^{(s+1)}$ is generated as follows:
  1. update U: sample $u^{(s+1)} \sim p_0(u|v^{(s)})$
  2. update V: sample $v^{(s+1)} \sim p_0(v|u^{(s+1)})$
- **Metropolis algorithm** (J_u and J_v are separate symmetric proposal distributions for U and V.)
  Propose and then accept or reject changes to one element at a time:
  1. update U:

- Sample $\mu^* \sim J_u(u|u^{(s)})$
- Compute the acceptance ratio $r = p_0(u^*, v^{(s)})/p_0(u^{(s)}, v^{(s)})$
- set $u^{(s+1)}$ to u^* or $u^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$.

2. update V:
   - Sample $v^* \sim J_v(v|v^{(s)})$
   - Compute the acceptance ratio $r = p_0(u^{(s+1)}, v^*)/p_0(u^{(s+1)}, v^{(s)})$
   - set $v^{(s+1)}$ to v^* or $v^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$.

- **Metropolis-Hastings algorithm**

  A Metropolis-Hastings algorithm for approximating $p_0(u, v)$ runs as follows:

1. update U:
   - Sample $\mu^* \sim J_u(u|u^{(s)}, v^{(s)})$
   - Compute the acceptance ratio

$$r = \frac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})} \times \frac{J_u(u^{(s)}|u^*, v^{(s)})}{J_u(u^*|u^{(s)}, v^{(s)})};$$

   - set $u^{(s+1)}$ to u^* or $u^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$.

2. update V:
   - Sample $v^* \sim J_v(v|u^{(s+1)}, v^{(s)})$
   - Compute the acceptance ratio

$$r = \frac{p_0(u^{(s+1)}, v^*)}{p_0(u^{(s+1)}, v^{(s)})} \times \frac{J_v(v^{(s)}|u^{(s+1)}, v^*)}{J_v(v^*|u^{(s+1)}, v^{(s)})}$$

- set $v^{(s+1)}$ to v^* or $v^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1 - r)$.

The proposal distributions J_u and J_v are not required to be symmetric.
$\frac{J_u(u^{(s)}|u^*,v^{(s)})}{J_u(u^*|u^{(s)},v^{(s)})}$, the ratio of the probability of generating the current value from the proposed to the probability of generating the proposed from the current. This can be viewed as a "correction factor": If a value u^* is much more likely to be proposed than the current value $u(s)$, then we must down-weight the probability of accepting u^* accordingly, otherwise the value u^* will be overrepresented in our sequence.

- **relationship and distinction:**
  - Metropolis algorithm: generate proposals from J_u and J_v and accepts them with some probability $\min(1, r)$.
  - Each step of the Gibbs sampler: generate a proposal from a full conditional distribution and then accepting it with probability 1.
  - Metropolis-Hastings algorithm: generalize both of these approaches by allowing arbitrary proposal distributions (which can be any distribution).

Metropolis algorithm $\sim$ Metropolis-Hastings algorithm: If J_u is symmetric, meaning that $J(u_a|u_b, v) = J(u_b|u_a, v)$ for all possible u_a, u_b and v, then the correction factor in the Metropolis-Hastings acceptance ratio is equal to 1 and the acceptance probability is the same as in the Metropolis algorithm.

Gibbs sampler $\sim$ Metropolis-Hastings: In Gibbs, the proposal distribution for U is the full conditional distribution of U given $V = \nu$. If we use the full conditionals as our proposal distributions in the Metropolis-Hastings, then $J_u(u^* \mid u^{(s)}, \nu^{(s)}) = p_0(u^* \mid \nu(s))$. The Metropolis-Hastings acceptance ratio is then

$$\begin{aligned} \mathrm{r} &= \frac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})} \times \frac{J_u(u^{(s)}|u^*, v^{(s)})}{J_u(u^*|u^{(s)}, v^{(s)})} = \frac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})} \frac{p_0(u^{(s)}|v^{(s)})}{p_0(u^*|v^{(s)})} \\ &= \frac{p_0(u^*|v^{(s)})p_0(v^{(s)})}{p_0(u^{(s)}|v^{(s)})p_0(v^{(s)})} \frac{p_0(u^{(s)}|v^{(s)})}{p_0(u^*|v^{(s)})} = \frac{p_0(v^{(s)})}{p_0(v^{(s)})} = 1, \end{aligned}$$

10.3.2 Why does the Metropolis-Hastings algorithm work?

A more general form of the Metropolis-Hastings algorithm: Given a current value $x^{(s)}$ of X,

1. Generate x^* from $J_s(x^*|x^{(s)})$
2. Compute the acceptance ratio
$$r = \frac{p_0(x^*)}{p_0(x^{(s)})} \times \frac{J_s(x^{(s)}|x^*)}{J_s(x^*|x^{(s)})};$$
3. Sample $\mu \sim \text{uniform}(0, 1)$. If $u < r$ set $x^{(s+1)} = x^*$, else set $x^{(s+1)} = x^{(s)}$

Note that the proposal distribution may also depend on the iteration number s. For example, the Metropolis-Hastings algorithm presented in the last section can be equivalently described by steps 1, 2 and 3 above by setting J_s to be equal to J_u for odd values of s and equal to J_v for even values. This makes the algorithm alternately update values of U and V.

The requirement of J_s:

- $J_s(x^* \mid x(s))$ does not depend on values in the sequence previous to $x(s)$.
- choose J_s so that the Markov chain is able to converge to the target distribution p0.

Theorem 10.1 *(Ergodic Theorem) If $\{x^{(1)}, x^{(2)}, \ldots\}$ is an irreducible, aperiodic and recurrent Markov chain, then there is a unique probability distribution π such that as $s \to \infty$,*

- $\Pr(x^{(s)} \in A) \longrightarrow \pi(A)$ *for any set A*;
- $\frac{1}{S}\sum g(x^{(s)}) \longrightarrow \int g(x)\pi(x)dx$.

The distribution π is called the stationary distribution of the Markov chain, because it has the following

property:

If $x(s) \sim \pi$, and $x(s+1)$ is generated from the Markov chain starting at $x(s)$, then $Pr(x(s+1) \in A) = \pi(A)$.

In other words, if you sample $x(s)$ from π and then generate $x(s+1)$ conditional on $x(s)$ from the Markov chain, then the unconditional distribution of $x(s+1)$ is π. Once you are sampling from the stationary distribution, you are always sampling from the stationary distribution.

What is left to show is that the stationary distribution π for a Metropolis-Hastings algorithm is equal to the distribution p_0 we wish to approximate.

"Proof" that $\pi(x) = p0(x)$

1. sampling $x^{(s)} = x_a$ from p_0
2. proposing $x^* = x_b$ from $J_s(x^*|x^{(s)})$
3. accepting $x^{(s+1)} = x_b$

$$\begin{aligned}\Pr(x^{(s)} = x_a, x^{(s+1)} = x_b) &= p_0(x_a) \times J_s(x_b|x_a) \times \frac{p_0(x_b)}{p_0(x_a)} \frac{J_s(x_a|x_b)}{J_s(x_b|x_a)} \\ &= p_0(x_b) J_s(x_a|x_b).\end{aligned}$$

$$\begin{aligned}\Pr(x^{(s+1)} = x) &= \sum_{x_a} \Pr(x^{(s+1)} = x, x^{(s)} = x_a) = \sum_{x_a} \Pr(x^{(s+1)} = x_a, x^{(s)} = x) \\ &= \Pr(x^{(s)} = x)\end{aligned}$$

$\Pr(x^{(s+1)} = x) = p_0(x)$ if $\Pr(x^{(s)} = x) = p_0(x)$.

10.4 Combining the Metropolis and Gibbs algorithms

The case: conditional distributions are available for some parameters but not for others.

$\Rightarrow$ combine Gibbs and Metropolis-type proposal distributions to generate a Markov chain to approximate the joint posterior distribution of all of the parameters.

10.4.1 A regression model with correlated errors

The ordinary regression model: $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim$ multivariate normal$(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

If the error terms are not independent, but temporally correlated

$\Rightarrow$ we must replace the covariance matrix $\sigma^2\mathbf{I}$ in the ordinary regression model with a matrix Σ that can represent positive correlation between sequential observations $\Rightarrow$ One simple and popular class of covariance matrices for temporally correlated data are those having first-order autoregressive structure:

$$\Sigma = \sigma^2\mathbf{C}_\rho = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \cdots \rho^{n-1} \\ \rho & 1 & \rho \cdots \rho^{n-2} \\ \rho^2 & \rho & 1 \\ \vdots & \vdots & \ddots \\ \rho^{n-1} & \rho^{n-2} & 1 \end{pmatrix}$$

Having observed $\boldsymbol{Y} = \boldsymbol{y}$, the parameters to estimate in this model include $\boldsymbol{\beta}$, σ^2 and ρ. Using the multivariate normal and inverse-gamma prior distributions for $\boldsymbol{\beta}$ and σ^2:

$$\begin{aligned}
\{\beta|\mathbf{X}, \boldsymbol{y}, \sigma^2, \rho\} &\sim \text{multivariate normal}(\beta_n, \Sigma_n), \text{where} \\
\sum_n &= (\mathbf{X}^T\mathbf{C}_\rho^{-1}\mathbf{X}/\sigma^2 + \Sigma_0^{-1})^{-1} \\
\boldsymbol{\beta}_n &= \Sigma_n(\mathbf{X}^T\mathbf{C}_\rho^{-1}y/\sigma^2 + \Sigma_0^{-1}\beta_0), \text{ and} \\
\{\sigma^2|\mathbf{X}, \boldsymbol{y}, \boldsymbol{\beta}, \rho\} &\sim \text{inverse-gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_\rho]/2), \text{where} \\
\text{SSR}_\rho &= (y - \mathbf{X}\beta)^T\mathbf{C}_\rho^{-1}(y - \mathbf{X}\beta).
\end{aligned} \tag{10.2}$$

- If $\boldsymbol{\beta}_0 = \mathbf{0}$ and Σ_0 has large diagonal entries, then $\boldsymbol{\beta}_n$ is very close to $(\mathbf{X}^T\mathbf{C}_\rho^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}_\rho^{-1}\boldsymbol{y}$. If ρ were known $\Rightarrow$ the generalized least squares (GLS) estimate of $\boldsymbol{\beta}$, a type of weighted least squares estimate that is used when the error terms are not independent and identically distributed.
- In such situations, both OLS and GLS provide unbiased estimates of $\boldsymbol{\beta}$ but GLS has a lower variance.
- Bayesian analysis using a model that accounts for the correlated errors provides parameter estimates that are similar to those of GLS, so for convenience we will refer to our analysis as "Bayesian GLS".

Given $\left\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}, \rho^{(s)}\right\}$, a Metropolis-Hastings algorithm to generate a new set of parameter values is as follows (use Metropolis algorithm to update ρ, use Gibbs to update $\boldsymbol{\beta}$ and σ^2):

1. Update $\boldsymbol{\beta}$: Sample $\boldsymbol{\beta}^{(s+1)} \sim$ multivarite normal $(\boldsymbol{\beta}_n, \Sigma_n)$, where $\boldsymbol{\beta}_n$ and Σ_n depend on $\sigma^{2(s)}$ and $\rho^{(s)}$
2. Update σ^2: Sample $\sigma^{2(s+1)} \sim$ inverse-gamma$([v_0 + n]/2, [v_0\sigma_0^2 + SSR_\rho]/2)$, where SSR_ρ depends on $\boldsymbol{\beta}^{(s+1)}$ and $\rho^{(s)}$
3. Update ρ:
   - Propose $\rho^* \sim$ uniform$(\rho^{(s)} - \delta, \rho^{(s)} + \delta)$ (a reflecting random walk which ensures that $0 < \rho < 1$). If $\rho^* < 0$ then reassign it to be $|\rho^*|$. If $\rho^* > 1$ reassign it to be $2 - \rho^*$
   - Compute the acceptance ratio

$$r = \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\sigma}^{2(s+1)}, \boldsymbol{\rho}^*)p(\boldsymbol{\rho}^*)}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\sigma}^{2(s+1)}, \boldsymbol{\rho}^{(s)})p(\boldsymbol{\rho}^{(s)})} = \frac{p(\boldsymbol{\beta}^{(s+1)}, \sigma^{2(s+1)}, \rho^*|\boldsymbol{y}, \mathbf{X})}{p(\boldsymbol{\beta}^{(s+1)}, \sigma^{2(s+1)}, \rho^{(s)}|\boldsymbol{y}, \mathbf{X})},$$

   second term: the ratio as given in the definition of the Metropolis algorithm; and sample $u \sim$ uniform$(0, 1)$. If $u < r$ set $\rho^{(s+1)} = \rho^*$, otherwise get $\rho^{(s+1)} = \rho^{(s)}$.

10.5 Discussion and further references

One technique (modifications and extensions of MCMC methods) that is broadly applicable is automatic, adaptive tuning of the proposal distribution in order to achieve good mixing.

Not all adaptive algorithms will result in chains that converge to the target distribution, but there are known conditions under which convergence is guaranteed.

11 Linear and generalized linear mixed effects models

- variation in the data was represented with a between-group sampling model for group-specific means
- a within-group sampling model to represent heterogeneity of observations within a group.
- extend the hierarchical model to describe how relationships between variables may differ between groups.
  - This can be done with a regression model to describe within-group variation, and a multivariate normal model to describe heterogeneity among regression coefficients across the groups
- estimation for hierarchical generalized linear models, which are hierarchical models that have a generalized linear regression model representing within-group heterogeneity.

11.1 A hierarchical regression model

- Use ordinary regression model to describe within-group heterogeneity of observations
- Use sampling model for the group-specific regression parameters to describe between-group heterogeneity.
- **Within-group sampling model:**

$$Y_{i,j} = \boldsymbol{\beta}_j^T \boldsymbol{x}_{i,j} + \epsilon_{i,j}, \{\epsilon_{i,j}\} \sim \text{i.i.d. normal}(0, \sigma^2), \tag{11.1}$$

where $\boldsymbol{x}_{i,j}$ is a $p \times 1$ vector of regressors for observation i in group j.
Expressing $Y_{1,j}, \ldots, Y_{n_j,j}$ as a vector $\boldsymbol{Y}_j$ and combining $\boldsymbol{x}_{1,j}, \ldots, \boldsymbol{x}_{n_j,j}$ into an $n_j \times p$ matrix $\mathbf{X}_j$. Then the within-group sampling model:

$$\boldsymbol{Y}_j \sim \text{multivariate normal}(\mathbf{X}_j \boldsymbol{\beta}_j, \sigma^2 \mathbf{I}),$$

with the group-specific data vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ being conditionally independent given $\beta_1, \ldots, \beta_m$ and σ^2

- **Between-group sampling model:**

  Describing the heterogeneity among the regression coefficients $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$

  If no prior information distinguishing the different groups $\Rightarrow$ model them as being exchangeable, or (roughly) equivalently, as being i.i.d. from some distribution representing the sampling variability across groups.

  The normal hierarchical regression model describes the across-group heterogeneity with a multivariate normal model, so that:

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \sim \text{i.i.d. multivariate normal}(\boldsymbol{\theta}, \Sigma). \tag{11.2}$$

Figure 3: the multivariate normal distribution for $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$ is not a prior distribution representing uncertainty about a fixed but unknown quantity. Rather, it is a sampling distribution representing heterogeneity among a collection of objects (likelihood ?). The values of $\boldsymbol{\theta}$ and Σ are fixed but unknown parameters to be estimated.

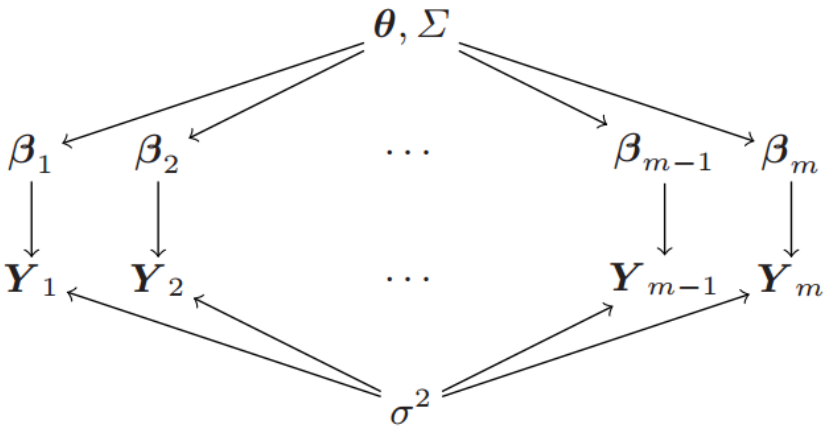


Figure 3: A graphical representation of the hierarchical normal regression model.

Hierarchical regression model; another name: linear mixed effects model, motivated by an alternative parameterization of Eqs 11.1 and 11.2.
We can rewrite the between-group sampling model as:

$$\boldsymbol{\beta}_j = \boldsymbol{\theta} + \boldsymbol{\gamma}_j$$
$$\boldsymbol{\gamma}_1 \ldots, \boldsymbol{\gamma}_m \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \Sigma).$$

- **Hierarchical regression model:**
  Plugging the above rewrite into within-group regression model gives

$$Y_{i,j} = \boldsymbol{\beta}_j^T \boldsymbol{x}_{i,j} + \epsilon_{i,j} = \boldsymbol{\theta}^T \boldsymbol{x}_{i,j} + \boldsymbol{\gamma}_j^T \boldsymbol{x}_{i,j} + \epsilon_{i,j}.$$

  In this parameterization $\boldsymbol{\theta}$ is referred to as a **fixed effect** as it is constant across groups, whereas $\gamma_1, \cdots, \gamma_m$ are called **random effects**, as they vary.
  Hierarchical regression model, another name: "mixed effects model", coming from the fact that the regression model contains both fixed and random effects. Although for our particular example the regressors corresponding to the fixed and random effects are the same, this does not have to be the case.
- **A more general model:**

$$Y_{i,j} = \boldsymbol{\theta}^T \boldsymbol{x}_{i,j} + \boldsymbol{\gamma}_j^T \boldsymbol{z}_{i,j} + \epsilon_{i,j},$$

  where $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_{i,j}$ could be vectors of different lengths which may or may not contain overlapping variables.
  - $\boldsymbol{x}_{i,j}$ contain group-specific regressors (constant across all observations in the same group).
  - $\boldsymbol{x}_{i,j}$ are not generally included in $\boldsymbol{z}_{i,j}$, as there would be no information in the data with which to estimate the corresponding group-specific regression coefficients.

Given a prior distribution for $(\boldsymbol{\theta}, \Sigma, \sigma^2)$ and having observed $\boldsymbol{Y}_1 = \boldsymbol{y}_1, \cdots, \boldsymbol{Y}_m = \boldsymbol{y}_m$, a Bayesian analysis proceeds by computing the posterior distribution:

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\theta}, \Sigma, \sigma^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_m, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m).$$

If semiconjugate prior distributions are used for $\boldsymbol{\theta}$, Σ and σ^2, then the posterior distribution can be approximated quite easily with Gibbs sampling. The classes of semiconjugate prior distributions for $\boldsymbol{\theta}$ and Σ are as in the multivariate normal model discussed in Chapter 7. The prior for σ^2 is the usual inverse-gamma distribution

$$\begin{aligned}
\boldsymbol{\theta} &\sim \text{multivariate normal}(\boldsymbol{\mu}_0, \Lambda_0) \\
\Sigma &\sim \text{inverse-Wishart}(\eta_0, \mathbf{S}_0^{-1}) \\
\sigma^2 &\sim \text{inverse-gamma}(\nu_0/2, \nu_0\sigma_0^2/2)
\end{aligned}$$

11.2 Full conditional distributions

Full conditional distributions $\Rightarrow$ iteratively sample from them to approximate the joint posterior distribution

11.2.1 Full conditional distributions of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$

Our hierarchical regression model shares information across groups via the parameters $\boldsymbol{\theta}$, Σ and σ^2. As a result, conditional on $\boldsymbol{\theta}$, Σ and σ^2 the regression coefficients $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$ are independent. Referring to the graph in Figure 3, from the perspective of a given $\boldsymbol{\beta}_j$ the model looks like an ordinary one-group regression problem where the prior mean and variance for $\boldsymbol{\beta}_j$ are $\boldsymbol{\theta}$ and Σ.

The results of Section 9.2.1 show that $\{\boldsymbol{\beta}_j | \boldsymbol{y}_j, \mathbf{X}_j, \boldsymbol{\theta}, \Sigma, \sigma^2\}$ has a multivariate normal distribution with

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{\beta}_j \mid \boldsymbol{y}_j, \mathbf{X}_j, \sigma^2, \theta, \Sigma] &= (\Sigma^{-1} + \mathbf{X}_j^T \mathbf{X}_j/\sigma^2)^{-1} \\
\mathrm{E}[\boldsymbol{\beta}_j \mid \boldsymbol{y}_j, \mathbf{X}_j, \sigma^2, \theta, \Sigma] &= (\Sigma^{-1} + \mathbf{X}_j^T \mathbf{X}_j/\sigma^2)^{-1}(\Sigma^{-1}\theta + \mathbf{X}_j^T \boldsymbol{y}_j/\sigma^2).
\end{aligned}$$

11.2.2 Full conditional distributions of θ, Σ

Sampling model for the $\boldsymbol{\beta}_j$'s: they are i.i.d. samples from a multivariate normal population with $\boldsymbol{\theta}$ and Σ. (C7: The full conditional distribution of a population mean is multivariate normal with expectation equal to a combination of the prior expectation and the sample mean, and precision equal to the sum of the prior and data precisions.) In the hierarchical regression model, given Σ and our sample of regression coefficients $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$, the full conditional distribution of $\boldsymbol{\theta}$ is as follows:

$$\{\boldsymbol{\theta} \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \Sigma\} \sim \text{multivariate normal}(\boldsymbol{\mu}_m, \Lambda_m)\ , \text{ where}$$
$$\Lambda_m = (\Lambda_0^{-1} + m\Sigma^{-1})^{-1}, \quad \boldsymbol{\mu}_m = \Lambda_m(\Lambda_0^{-1}\boldsymbol{\mu}_0 + m\Sigma^{-1}\bar{\boldsymbol{\beta}}), \quad \bar{\boldsymbol{\beta}} \text{ is the vector average } 1/m \sum \boldsymbol{\beta}_j.$$

C7: The full conditional distribution of a covariance matrix is an inverse-Wishart distribution, with sum of squares matrix equal to the prior sum of squares $\mathbf{S}_0$ plus the sum of squares from the sample:

$$\{\Sigma \mid \boldsymbol{\theta}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m\} \sim \text{inverse-Wishart } (\eta_0 + m, [\mathbf{S}_0 + \mathbf{S}_\theta]^{-1}), \text{where } \mathbf{S}_\theta = \sum_{j=1}^{m} (\boldsymbol{\beta}_j - \boldsymbol{\theta})(\boldsymbol{\beta}_j - \boldsymbol{\theta})^T.$$

Note that $\mathbf{S}_\theta$ depends on $\boldsymbol{\theta}$ and so must be recomputed each time $\boldsymbol{\theta}$ is updated in the Markov chain.

11.2.3 Full condition distribution of σ^2

Parameter σ^2 represents the error variance, assumed to be **common** across all groups. As such, conditional on $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$, the data provide information about σ^2 via the sum of squared residuals from each group:

$$\sigma^2 \sim \text{inverse-gamma } ([\nu_0 + \sum n_j]/2, [\nu_0\sigma_0^2 + \text{SSR}]/2), \text{where SSR} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{i,j} - \boldsymbol{\beta}_j^T \boldsymbol{x}_{i,j})^2.$$

Note that SSR depends on the value of $\boldsymbol{\beta}_j$, and so SSR must be recomputed in each scan of the Gibbs sampler before σ^2 is updated.

11.3 Generalized linear mixed effects models

A generalized linear mixed effects model combines aspects of linear mixed effects models with those of generalized linear models described in Chapter 10.

Such models are useful when we have a hierarchical data structure but the normal model for the within-group variation is not appropriate.

- Eg: if the variable Y were binary or a count, then more appropriate models for within-group variation would be logistic or Poisson regression models, respectively.

A basic generalized linear mixed model is as follows:

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \sim \text{i.i.d. multivariate normal } (\boldsymbol{\theta}, \Sigma)$$

$$p(\boldsymbol{y}_j \mid \mathbf{X}_j, \boldsymbol{\beta}_j, \gamma) = \prod_{i=1}^{n_j} p(y_{i,j} \mid \boldsymbol{\beta}_j^T \boldsymbol{x}_{i,j}, \gamma),$$

with observations from different groups also being conditionally independent.

In this formulation $p(y \mid \boldsymbol{\beta}^T \boldsymbol{x}, \gamma)$ is a density whose mean depends on $\boldsymbol{\beta}^T \boldsymbol{x}$, and γ is an additional parameter often representing variance or scale. For example:

- In the normal model $p(y \mid \boldsymbol{\beta}^T \boldsymbol{x}, \gamma) = \text{dnorm}(y, \boldsymbol{\beta}^T \boldsymbol{x}, \gamma^{1/2})$ where γ represents the variance.
- In the Poisson model $p(y \mid \boldsymbol{\beta}^T \boldsymbol{x}) = \text{dpois}(\exp\{\boldsymbol{\beta}^T \boldsymbol{x}\})$, and there is no γ parameter.

11.3.1 A Metropolis-Gibbs algorithm for posterior approximation

- Estimation for the linear mixed effects model was straightforward because the full conditional distribution of each parameter was standard, allowing for the easy implementation of a Gibbs sampling algorithm.
- But for nonnormal generalized linear mixed models, only $\boldsymbol{\theta}$ and Σ have standard full condi. distributions. Thus, use a Metropolis-Hastings algorithm to approximate the posterior distribution of the parameters, using a combination of Gibbs steps for updating $(\boldsymbol{\theta}, \Sigma)$ with a Metropolis step for updating each $\boldsymbol{\beta}_j$.
  - We assume there is no γ parameter. If there is such a parameter, it can be updated using a Gibbs step if a full conditional distribution is available, and a Metropolis step if not.

1. **Gibbs steps for $\boldsymbol{\theta}, \Sigma$**
   As in the linear mixed effects model, the full conditional distributions of $\boldsymbol{\theta}$ and Σ depend only on $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$. Thus, the form of $p(y \mid \boldsymbol{\beta}^T \boldsymbol{x})$ has no effect on the full conditional distributions of $\boldsymbol{\theta}$ and Σ. Whether $p(y \mid \boldsymbol{\beta}^T \boldsymbol{x})$ is a normal/Poisson/some other generalized linear model, the full conditional distributions of $\boldsymbol{\theta}$ and Σ will be the multivariate normal and inverse-Wishart distributions in Section 11.2.
2. **Metropolis step for $\boldsymbol{\beta}_j$**
   Updating $\boldsymbol{\beta}_j$ in a Markov chain: proceed by proposing a new value of $\boldsymbol{\beta}_j^*$ based on the current parameter values and accepting or rejecting it with appropriate probability. Standard proposal distribution: multivariate normal distribution with mean (current value $\boldsymbol{\beta}_j^{(s)}$) and proposal variance $V_j^{(s)}$. Metropolis step:
   (a). Sample $\boldsymbol{\beta}_j^* \sim$ multivariate normal$(\boldsymbol{\beta}_j^{(s)}, V_j^{(s)})$
   (b). Compute the acceptance ratio $r = \dfrac{p(\boldsymbol{y}_j|\mathbf{X}_j, \boldsymbol{\beta}_j^*)p(\boldsymbol{\beta}_j^*|\boldsymbol{\theta}^{(s)}, \Sigma^{(s)})}{p(\boldsymbol{y}_j|\mathbf{X}_j, \boldsymbol{\beta}_j^{(s)})p(\boldsymbol{\beta}_j^{(s)}|\boldsymbol{\theta}^{(s)}, \Sigma^{(s)})}$.
   (c). Sample $u \sim$uniform$(0, 1)$. Set $\boldsymbol{\beta}_j^{(s+1)}$ to $\boldsymbol{\beta}_j^*$ if $u < r$ and to $\boldsymbol{\beta}_j^{(s)}$ if $u > r$

3. **A Metropolis-Hastings approximation algorithm**
   Putting these steps together $\Rightarrow$ the following Metropolis-Hastings algorithm for approximating $p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\theta}, \Sigma|\mathbf{X}$
   Given current values at scan s of the Markov chain, we obtain new values as follows:
   (a). Sample $\boldsymbol{\theta}^{(s+1)}$ from its full conditional distribution
   (b). Sample $\Sigma^{(s+1)}$ from its full conditional distribution
   (c). For each $j \in \{1, \ldots, m\}$,
   - propose a new value $\boldsymbol{\beta}_j^*$;
   - set $\boldsymbol{\beta}_j^{(s+1)}$ equal to $\boldsymbol{\beta}_j^*$ or $\boldsymbol{\beta}_j^{(s)}$ with the appropriate probability.

11.4 Posterior analysis of the math score data

The math score data described in Section 8.4: including math scores of 10th grade children from 100 different large urban public high schools.

- In Chapter 8: estimated school-specific expected math scores, as well as how these expected values varied from school to school.
- Suppose we are interested in examining the relationship between math score and another variable, socioeconomic status (SES), which was calculated from parental income and education levels for each student in the dataset.

In Chapter 8 we quantified the between-school heterogeneity in expected math score with a hierarchical model. It seems possible that the relationship between math score and SES might vary from school to school as well. A quick and easy way to assess this possibility is to fit a linear regression model of math score as a function of SES for each of the 100 schools in the dataset. To make the parameters more interpretable we will

center the SES scores within each school separately, so that the sample average SES score within each school is zero. As a result, the intercept of the regression line can be interpreted as the school-level average math score.

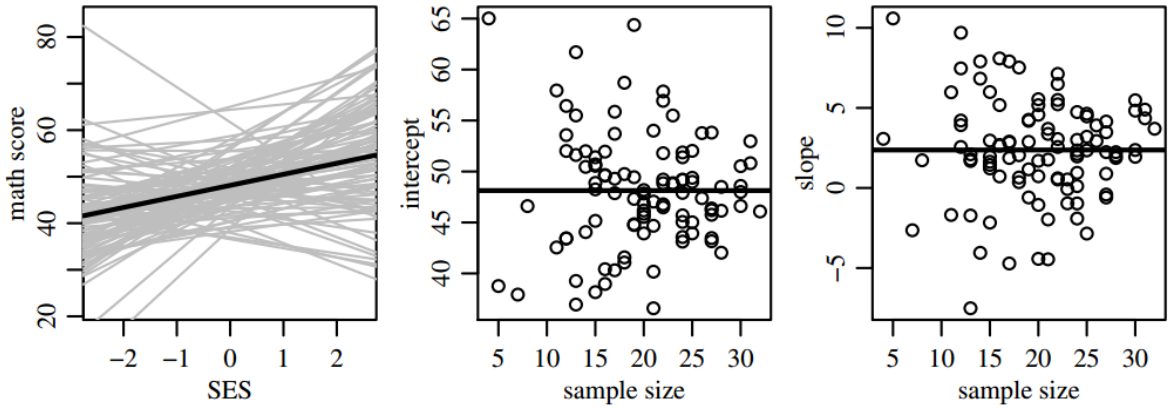


Figure 4: Least squares regression lines for math score data, and plots of estimates versus group sample size.

The first panel of Figure 4: least squares estimates of the regression lines for the 100 schools, along with an average of these lines in black. A large majority show an increase in expected math score with increasing SES, although a few show a negative relationship. The second and third panels of the figure relate the least squares estimates to sample size.

Notice that schools with the highest sample sizes have regression coefficients that are generally close to the average, whereas schools with extreme coefficients are generally those with low sample sizes. This phenomenon is reminiscent of what we discussed in Section 8.4: The smaller the sample size for the group, the more probable that unrepresentative data are sampled and an extreme least squares estimate is produced.As in Chapter 8, our remedy to this problem will be to stabilize the estimates for small sample size schools by sharing information across groups, using a hierarchical model.

To analyze the math score data, use a prior distribution that is similar in spirit to the unit information priors that were discussed in Chapter 9.

- Take $\boldsymbol{\mu}_0$, the prior expectation of $\boldsymbol{\theta}$, to be equal to the average of the ordinary least squares regression estimates and the prior variance Λ_0 to be their sample covariance.
  Such a prior distribution represents the information of someone with unbiased but weak prior information. For example, a 95% prior confidence interval for the slope parameter θ_2 under this prior is (3.86,8.60), which is quite a large range considering what the extremes of the interval imply in terms of average change in score per unit change in SES score.
- Take the prior sum of squares matrix $\mathbf{S}_0$ to be equal to the covariance of the least squares estimate, but take the prior degrees of freedom η_0 to be $p + 2 = 4$, so that the prior distribution of Σ is reasonably diffuse but has an expectation equal to the sample covariance of the least squares estimates.
- Take σ_0^2 to be the average of the within-group sample variance but set $\nu_0 = 1$.

Running a Gibbs sampler for 10,000 scans and saving every 10th scan produces a sequence of 1,000 values for each parameter, each sequence having a fairly low autocorrelation. As usual, we can use these simulated values to make Monte Carlo approximations to various posterior quantities of interest.

For example, the first plot in Figure 11.3 shows the posterior distribution of θ_2, the expected within-school slope parameter. A 95% quantile-based posterior confidence interval for this parameter is (1.83, 2.96), which, compared to our prior interval of (-3.86, 8.60), indicates a strong alteration in our information about θ_2.

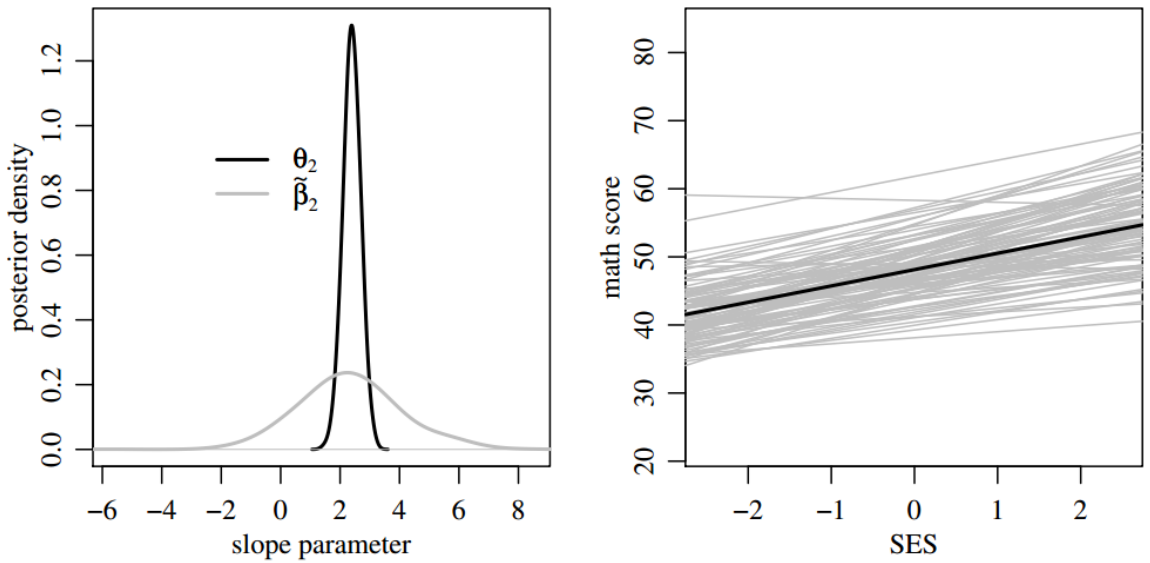


Figure 5: Relationship between SES and math score. The first panel: the posterior density of the expected slope θ_2 of a randomly sampled school, and the posterior predictive distribution of a randomly sampled slope. The second panel: posterior expectations of the 100 school-specific regression lines, with black average line.

The fact that θ_2 is extremely unlikely to be negative only indicates that the population averages of school-level slopes is positive. It does not indicate that any given within-school slope cannot be negative. To clarify this distinction, the posterior predictive distribution of $\widetilde{\beta}_2$, the slope for a to-be-sampled school, is plotted in the same figure. Samples from this distribution can be generated by sampling a value $\widetilde{\beta}_2(s)$ from a multivariate normal$(\theta(s), \Sigma(s))$ distribution for each scan s of the Gibbs sampler. Notice that this posterior predictive distribution is much more spread out than the posterior distribution of θ_2, reflecting the heterogeneity in slopes across schools. Using the Monte Carlo approximation, we have $\Pr(\widetilde{\beta}_2 < 0 \mid \boldsymbol{y}_1, |, \boldsymbol{y}_m, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_m) \approx 0.07$, which is small but not negligible.

The second panel in Figure 5 plots posterior expectations of the 100 school-specific regression lines, with the line given by the posterior mean of θ in black. Comparing this to the first panel of Figure 4 indicates how the hierarchical model is able to share information across groups, shrinking extreme regression lines towards the across-group average. Particularly, hardly any of the slopes are negative when we share information across groups.

11.5 Discussion and further references

Posterior approximation via MCMC for hierarchical models can suffer from poor mixing. One reason for this is that many of the parameters in the model are highly correlated, and generating them one at a time in the Gibbs sampler can lead to a high degree of autocorrelation.

For example, $\boldsymbol{\theta}$ and the $\boldsymbol{\beta}_j$'s are positively correlated, and so an extreme value of $\boldsymbol{\theta}$ at one iteration can lead to extreme values of the $\boldsymbol{\beta}_j$'s when they get updated, especially if the amount of within-group data is low. This in turn leads to an extreme value of $\boldsymbol{\theta}$ at the next iteration.

12 Latent variable methods for ordinal data

Motivation

- Many datasets include variables whose distributions cannot be represented by typical distributions.
- Such variables are binned into ordered categories, the number of which may vary from survey to survey.

In such situations, interest often lies in the associations between the variables: Is the relationship between two variables positive, negative or zero? What happens if we "account" for a third variable?

- For normally distributed data: these types of questions can be addressed with the multivariate normal and linear regression models of Chapters 7 and 9.
- For nonnormal data: extends these models to this situation by expressing non-normal random variables as functions of unobserved, "latent" normally distributed random variables.
  $\Rightarrow$ Multivariate normal and linear regression models then can be applied to the latent data.

12.1 Ordered probit regression and the rank likelihood

DEG Example

The relationship between the educational attainment and number of children of individuals in a population. Additionally, an individual's educational attainment may be influenced by their parent's education level.

Data: DEG_i: the highest degree obtained by individual i, CHILD_i: their number of children, PDEG_i: the binary indicator of whether or not either parent of i obtained a college degree.

First, investigate the relationship between the variables with a linear regression model:

$$\text{DEG}_i = \beta_1 + \beta_2 \times \text{CHILD}_i + \beta_3 \times \text{PDEG}_i + \beta_4 \times \text{CHILD}_i \times \text{PDEG}_i + \epsilon_i,$$

where we assume that $\epsilon_1, \cdots, \epsilon_n \sim$ i.i.d. normal$(0, \sigma^2)$.

However, such a model would be inappropriate for a couple of reasons:

1. Since the variable DEG takes on only a small set of discrete values, the normality assumption of the residuals will certainly be violated.
2. The regression model imposes a numerical scale to the data that is not really present: A bachelor's degree is not "twice as much" as a high school degree

Variables for which there is a logical ordering of the sample space are known as ordinal variables. The discrete variables DEG and CHILD are ordinal variables, as are "continuous" variables like height or weight. However, CHILD, height and weight are measured on meaningful numerical scales, whereas DEG is not.

DEG is ordinal but not numeric, whereas CHILD is ordinal, numeric and discrete. Variables like height or weight are ordinal, numeric and continuous.

- The term "ordinal": variable for which there is a logical ordering of the sample space.
- The term "numeric": variables that have meaningful numerical scales, and "continuous" if a variable can have a value that is (roughly) any real number in an interval.

12.1.1 Probit regression

It is natural to think of many ordinal, non-numeric variables as arising from some underlying numeric process. Eg: the amount of effort a person puts into formal education may lie within a continuum, but a survey may only record a rough, categorized version of this variable, such as DEG.

This idea motivates a modeling technique known as **ordered probit regression** (model): relate a variable Y to a vector of predictors $\boldsymbol{x}$ via a regression in terms of a latent variable Z. More precisely, the model is

$$\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. normal}(0, 1) \tag{12.1}$$

$$Z_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i \tag{12.2}$$

$$Y_i = g(Z_i), \tag{12.3}$$

where $\boldsymbol{\beta}$ and g are unknown parameters.

- The variance of $\epsilon_1, \cdots, \epsilon_n$ is one, because the scale of the distribution of Y can already be represented by g, as g is allowed to be any non-decreasing function.
- $\boldsymbol{\beta}$: regression coefficients, describing the relationship between the explanatory variables and the unobserved latent variable Z.
- g: non-decreasing function, which relates the value of Z to the observed variable $Y \Rightarrow$ the sign of a regression coefficient β_j indicates whether Y is increasing or decreasing in x_j. Besides, g can represent the location of the distribution of Y, and so we do not need to include an intercept term in the model.

If the sample space for Y takes on K values, say $\{1, ..., K\}$, then the function g can be described with only $K-1$ ordered parameters $g_1 < g_2 < \cdots < g_{K-1}$ as follows:

$$\begin{aligned} y = g(z) &= 1 \quad \text{if } -\infty = g_0 < z < g_1 \\ &= 2 \quad \text{if } g_1 < z < g_2 \\ &\vdots \\ &= K \quad \text{if } g_{K-1} < z < g_K = \infty. \end{aligned} \tag{12.4}$$

$\{g_1, g_2, \cdots, g_{K-1}\}$, maybe "thresholds", so moving z past a threshold moves y into the next highest category.

Unknown parameters: β and $g_1, \cdots, g_{K-1}$. Use normal prior distributions for these, the joint posterior distribution of $\{\boldsymbol{\beta}, g_1, \cdots, g_{K-1}, Z_1, \cdots, Z_n\}$ given $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \cdots, y_n)$ can be approximated by Gibbs.

1. **Full conditional distribution of $\boldsymbol{\beta}$**
   Given $\boldsymbol{Y} = \boldsymbol{y}$, $\boldsymbol{Z} = \boldsymbol{z}$, and $\boldsymbol{g} = (g_1, \cdots, g_{K-1})$, the full conditional distribution of $\boldsymbol{\beta}$ depends only on $\boldsymbol{z}$ and satisfies $p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{g}) \propto p(\boldsymbol{\beta}) p(\boldsymbol{z} \mid \boldsymbol{\beta})$.
   Just as in ordinary regression, a multivariate normal prior distribution for $\boldsymbol{\beta}$ gives a multivariate normal posterior distribution. For example, if we use $\boldsymbol{\beta} \sim$ multivariate normal$(\mathbf{0}, n(\mathbf{X}^T\mathbf{X})^{-1})$, then $p(\boldsymbol{\beta} \mid \boldsymbol{z})$ is multivariate normal with

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{g}) \propto p(\boldsymbol{\beta}) p(\boldsymbol{z} \mid \boldsymbol{\beta})$$

$$\text{Var}[\boldsymbol{\beta} \mid \boldsymbol{z}] = \frac{n}{n+1}(\mathbf{X}^T\mathbf{X})^{-1}, \text{ and}$$

$$\text{E}[\boldsymbol{\beta} \mid \boldsymbol{z}] = \frac{n}{n+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{z}.$$

2. **Full conditional distribution of $\boldsymbol{Z}$**
   Under the sampling model, the conditional distribution of Z_i given $\boldsymbol{\beta}$ is $Z_i \sim$normal$(\boldsymbol{\beta}^T\boldsymbol{x}_i, 1)$.
   Given g, observing $Y_i = y_i$ tells us that Z_i must lie in the interval (g_{y_i-1}, g_{y_i}). Letting $a = g_{y_i-1}$ and $b = g_{y_i}$, the full conditional distribution of Z_i given $\{\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{g}\}$ is:

$$p(z_i \mid \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{g}) \propto \text{dnorm}(z_i, \boldsymbol{\beta}^T\boldsymbol{x}_i, 1) \times \delta_{(a,b)}(z_i).$$

   This is the density of a constrained normal distribution. To sample a value x from a normal(μ, σ^2) distribution constrained to the interval (a, b), we perform the following two steps:
   (a). sample $u \sim$ uniform$(\Phi[(a-\mu)/\sigma], \Phi[(b-\mu)/\sigma])$
   (b). $x = \mu + \sigma\Phi^{-1}(u)$
   where Φ and Φ^{-1} are the cdf and inverse-cdf of the standard normal distribution.
3. **Full conditional distribution of $\boldsymbol{g}$**
   Suppose the prior distribution for $\boldsymbol{g}$ is some arbitrary density $p(\boldsymbol{g})$.
   Given $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{Z} = \boldsymbol{z}$, we know from Eq 12.4 that g_k must be higher than all z_i's for which $y_i = k$ and lower than all z_i's for which $y_i = k+1$.
   Letting $a_k = \max\{z_i : y_i = k\}$ and $b_k = \min\{z_i : y_i = k+1\}$. The full conditional distribution of $\boldsymbol{g}$ is then proportional to $p(\boldsymbol{g})$ but constrained to the set $\{\boldsymbol{g} : a_k < g_k < b_k\}$. For example, if $p(\boldsymbol{g})$ is proportional to the product $\prod_{k=1}^{K-1} \text{dnorm}(g_k, \mu_k, \sigma_k)$ but constrained so that $g_1 < \cdots < g_{K-1}$, then the full conditional density of g_k is a normal (μ_k, σ_k^2) density constrained to the interval (a_k, b_k).

12.1.2 Transformation models and the rank likelihood

DEG Example (cont.): Let Y_i be DEGi, $\boldsymbol{x}_i = (\text{CHILD}_i, \text{PDEG}_i, \text{CHILDi} \times \text{PDEG}_i)$.

- We need to specify a prior distribution for $\boldsymbol{\beta}$ and the transformation $g(z)$, as specified by the vector g of $K-1$ threshold parameters.
  - While simple default prior distributions for $\boldsymbol{\beta}$ exist (such as Zellner's g-prior), the same is not true for $\boldsymbol{g}$. Coming up with a prior distribution for $\boldsymbol{g}$ that represents actual prior information seems like a difficult task. Of course, this task is much harder if the number of categories K is large.
- An alternative approach to estimating $\boldsymbol{\beta}$ that does not require us to estimate the function $g(z)$:
  - If the Z_i's were observed directly: ignore Eq.12.3 of the model, left with an ordinary regression problem without having to estimate the transformation $g(z)$.
  - We do not observe the Z_i's directly, but there is information in the data about the Z_i's that does not require us to specify $g(z)$: Since we know that g is non-decreasing, we do know something about the order of the Z_i's. For example, if our observed data are such that $y_1 > y_2$, then since $y_i = g(Z_i)$, we know that $g(Z_1) > g(Z_2)$, this means that we know $Z_1 > Z_2$. In other words, having observed $\boldsymbol{Y} = \boldsymbol{y}$, we know that the Z_i's must lie in the set

$$R(\boldsymbol{y}) = \{z \in \mathbb{R}^n : z_{i_1} < z_{i_2} \quad \text{if} \quad y_{i_1} < y_{i_2}\}.$$

  - Since the distribution of the Z_i's does not depend on g, the probability that $\boldsymbol{Z} \in R(\boldsymbol{y})$ for a given $\boldsymbol{y}$ also does not depend on the unknown function g. Thus we base our posterior inference on the knowledge that $\boldsymbol{Z} \in R(\boldsymbol{y})$. Our posterior distribution for $\boldsymbol{\beta}$ in this case is given by

$$p(\boldsymbol{\beta} \mid \boldsymbol{Z} \in R(\boldsymbol{y})) \propto p(\boldsymbol{\beta}) \times \Pr(\boldsymbol{Z} \in R(\boldsymbol{y}) \mid \boldsymbol{\beta}) = p(\boldsymbol{\beta}) \times \int_{R(\boldsymbol{y})} \prod_{i=1}^{n} \text{dnorm}(z_i, \boldsymbol{\beta}^T \boldsymbol{x}_i, 1) dz_i.$$

  As a function of $\boldsymbol{\beta}$, the probability $\Pr(\boldsymbol{Z} \in R(\boldsymbol{y}) \mid \boldsymbol{\beta})$ is known as the rank likelihood.

1. For continuous data, it contains the same information about $\boldsymbol{y}$ as knowing the ranks of $\{y_1, \cdots, y_n\}$, i.e. which one has the highest value, which one has the second highest value, etc.
2. For discrete data, observing $R(\boldsymbol{y})$ is not exactly the same as knowing the ranks
3. But for simplicity we will still refer to $\Pr(\boldsymbol{Z} \in R(\boldsymbol{y}) \mid \boldsymbol{\beta})$ as the rank likelihood, whether or not Y is discrete or continuous.

- **Important:** for any ordinal outcome variable Y (non-numeric, numeric, discrete or continuous), information about $\boldsymbol{\beta}$ can be obtained from $\Pr(\boldsymbol{Z} \in R(\boldsymbol{y}) \mid \boldsymbol{\beta})$ without having to specify $g(z)$.
  - For any given $\boldsymbol{\beta}$ the value of $\Pr(\boldsymbol{Z} \in R(\boldsymbol{y}) \mid \boldsymbol{\beta})$ involves a complicated integral.
  - However, by estimating $\boldsymbol{Z}$ simultaneously with $\boldsymbol{\beta}$ we can obtain an estimate of $\boldsymbol{\beta}$ without ever having to numerically compute $\Pr(\boldsymbol{Z} \in R(\boldsymbol{y}) \mid \boldsymbol{\beta})$.

    The joint posterior distribution of $\{\boldsymbol{\beta}, \boldsymbol{Z}\}$ can be approximated by using Gibbs sampling (sampling from full conditional distributions).

    1. The full conditional distribution of $\boldsymbol{\beta}$:
       Given a current value $\boldsymbol{z}$ of $\boldsymbol{Z}$, the full conditional density $p(\boldsymbol{\beta} \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Z} \in R(\boldsymbol{y}))$ reduces to $p(\boldsymbol{\beta} \mid \boldsymbol{Z} = \boldsymbol{z})$ because knowing the value of $\boldsymbol{Z}$ is more informative than knowing just that $\boldsymbol{Z}$ lies in the set $R(\boldsymbol{y})$.
    2. A multivariate normal prior distribution for $\boldsymbol{\beta}$:
       results in a multivariate normal full conditional distribution, as before.
    3. The full conditional distributions of the Z_i's:
       Consider the full conditional distribution of Z_i given $\{\boldsymbol{\beta}, \boldsymbol{Z} \in R(\boldsymbol{y}), \boldsymbol{z}_{-i}\}$, where $\boldsymbol{z}_{-i}$ denotes the values of all of the Z's except Z_i. Conditional on $\boldsymbol{\beta}$, Z_i is normal$(\boldsymbol{\beta}^T \boldsymbol{x}_i, 1)$. Conditional on $\{\boldsymbol{\beta}, Z \in R(\boldsymbol{y}), \boldsymbol{z}_{-i}\}$, the density of Z_i is proportional to a normal density but constrained by the fact that $\boldsymbol{Z} \in R(\boldsymbol{y})$. Let's recall the nature of this constraint: $y_i < y_j$ implies $Z_i < Z_j$, and $y_i > y_j$

implies $Z_i > Z_j$. This means that Z_i must lie in the following interval:

$$\max\{z_j : y_j < y_i\} < Z_i < \min\{z_j : y_i < y_j\}.$$

Letting a and b denote the numerical values of the lower and upper endpoints of this interval, the full conditional distribution of Z_i is then

$$p(z_i \mid \boldsymbol{\beta}, \boldsymbol{Z} \in \boldsymbol{R}(\boldsymbol{y}), \boldsymbol{z}_{-i}) \propto \text{dnorm}(z_i, \boldsymbol{\beta}^T \boldsymbol{x}_i, 1) \times \delta_{(a,b)}(z_i).$$

This full conditional distribution is exactly the same as that of Z_i in the ordered probit model, except that now the constraints on Z_i are determined directly by the current value $\boldsymbol{Z}_{-i}$, instead of on the threshold variables r. As such, sampling from this full conditional distribution is very similar to sampling from the analogous distribution in the probit regression model.

DEG Example (cont.): For the educational attainment data the posterior distribution of $\boldsymbol{\beta}$ based on the rank likelihood is very similar to the one based on the full ordered probit model.

The full ordered probit model and the rank likelihood

In general, if K is small and n is large, we expect the two methods to behave similarly.

- The rank likelihood approach is applicable to a wider array of datasets since with this approach, Y is allowed to be any type of ordinal variable, discrete or continuous.
- Drawback of the rank likelihood: it does not provide inference about $g(z)$ describing the relationship between latent and observed variables. If this parameter is of interest, the rank likelihood is not appropriate; but if interest lies only in $\boldsymbol{\beta}$, this model provides a simple alternative to the ordered probit model.

12.2 The Gaussian copula model

- Above regression model limiting: only describes the conditional distribution of 1 variable given the others.
- We may be interested in the relationships among all of the variables in a dataset.
  - If the variables were approximately jointly normally distributed, or at least were all measured on a meaningful numerical scale $\Rightarrow$ describe relationships among variables with the sample covariance matrix or a multivariate normal model.
  - However, such a model is inappropriate for nonnumeric ordinal variables like INC, DEG and PDEG.
- To accommodate variables such as these we can extend the ordered probit model above to a latent, multivariate normal model that is appropriate for all types of ordinal data, both numeric and non-numeric.

Let $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ be i.i.d. random samples from a p-variate population, the latent normal model is

$$\begin{aligned} \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \Psi) \\ Y_{i,j} &= g_j(Z_{i,j}), \end{aligned} \tag{12.5}$$

where $g_1, \ldots, g_p$ are non-decreasing functions, Ψ is a correlation matrix, having diagonal entries equal to 1.

The matrix Ψ represents the joint dependencies among the variables, and the functions $g_1, \ldots, g_p$ represent their marginal distributions. To see how the g_j's represent the margins, let's calculate the marginal cdf $F_j(y)$ of a continuous random variable $Y_{i,j}$ under the model given by Eq 12.5.

Recalling the definition of the cdf, we have

$$\begin{aligned} F_j(y) &= \Pr(Y_{i,j} \leq y) \\ &= \Pr(g_j(Z_{i,j}) \leq y) \text{ , since } Y_{i,j} = g_j(Z_{i,j}) \\ &= \Pr(Z_{i,j} \leq g_j^{-1}(y)) \\ &= \Phi(g_j^{-1}(y)), \end{aligned}$$

where $\Phi(z)$ is the cdf of the standard normal distribution. The last line holds because the diagonal entries of Ψ are all equal to 1, and so the marginal distribution of each $Z_{i,j}$ is a standard normal distribution with cdf $\Phi(z)$.

$F_j(y) = \Phi(g_j^{-1}(y)) \Rightarrow$ Marginal distributions of the Y_j's are fully determined by the g_j's, not the matrix Ψ.

- A model having separate parameters for the univariate marginal distributions and the multivariate dependencies is generally called a *copula model*.
- The model given by Eq 12.5, where the dependence is described by a multivariate normal distribution, is called the *multivariate normal copula model*.
- The term "copula" refers to the method of "coupling" a model for multivariate dependence (such as the multivariate normal distribution) to a model for the marginal distributions of the data.

As shown above, a copula model separates the parameters for the dependencies among the variables Ψ from the parameters describing their univariate marginal distributions $g_1, \ldots, g_p$. This separation comes in handy if we are primarily interested in the dependencies among variables and not the univariate scales on which they were measured. In this case, $g_1, \ldots, g_p$ functions are nuisance parameters and Ψ is parameter interested.

Using an extension of the rank likelihood described in the previous section, we can obtain a posterior distribution for Ψ without having to estimate or specify prior distributions for the nuisance parameters $g_1, \ldots, g_p$.

12.2.1 Rank likelihood for copula estimation

- **The unknown parameters in the copula model:** the matrix Ψ and the nondecreasing functions $g_1, \ldots, g_p$.
- Bayesian inference for all of these parameters would require that we specify a prior for Ψ as well as p prior distributions over the complicated space of arbitrary non-decreasing functions.
  - If not interested in $g_1, \ldots, g_p \Rightarrow$ Use a version of the rank likelihood, quantifies information about $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ without having to specify these nuisance parameters.

  Each g_j: non-decreasing $\Rightarrow$ observe $n \times p$ data matrix $\mathbf{Y}$: the matrix of latent variables $\mathbf{Z}$ lie in set:

$$R(\mathbf{Y}) = \{\mathbf{Z} : z_{i_1,j} < z_{i_2,j} \text{if } y_{i_1,j} < y_{i_2,j}\}. \tag{12.6}$$

The probability of this event, $\Pr(\mathbf{Z} \in R(\mathbf{Y}) \mid \Psi)$, does not depend on $g_1, \ldots, g_p$. As a function of Ψ, $\Pr(\mathbf{Z} \in R(\mathbf{Y}) \mid \Psi)$ is called the rank likelihood for the multivariate normal copula model.

Compute the likelihood for a given Ψ: difficult, but as in Sec 12.1.2 we make an MCMC approximation to $p(\Psi, \mathbf{Z} \mid \mathbf{Z} \in R(\mathbf{Y}))$ using Gibbs sampling, provided we use a prior for Ψ based on the IW distribution.

Full conditional distribution of Ψ (A parameter-expanded prior distribution for Ψ)

Unfortunately there is no simple conjugate class of prior distributions for our correlation matrix Ψ.

As an alternative, let's consider altering Eq.12.5(1) to be

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \Sigma),$$

where Σ is an arbitrary covariance matrix, not restricted to be a correlation matrix like Ψ.

- In this case, a natural prior distribution for Σ: an inverse-Wishart distribution, which would give an inverse-Wishart full conditional distribution and thus make posterior inference available via Gibbs sampling.
- However, careful inspection of the rank likelihood indicates that it does not provide us with a complete estimate of Σ.
  - Specifically, the rank likelihood contains only information about the relative ordering among the $Z_{i,j}$'s, and no information about their scale. For example, if $Z_{1,j}$ and $Z_{2,j}$ are two i.i.d. samples from a normal$(0, \sigma_j^2)$ distribution, then the probability that $Z_{1,j} < Z_{2,j}$ does not depend on σ_j^2.

  For this reason we say that the diagonal entries of Σ are non-identifiable in this model, meaning that the rank likelihood provides no information about what the diagonal should be.

In a Bayesian analysis, the posterior distribution of any non-identifiable parameter is determined by the prior distribution, and so in some sense the posterior distribution of such a parameter is not of interest.

However, to each covariance Σ there corresponds a unique correlation matrix Ψ, obtained by the function:

$$\Psi = h(\Sigma) = \{\sigma_{i,j}/\sqrt{\sigma_i^2 \sigma_j^2}\}.$$

The value of Ψ is identifiable from the rank likelihood, and so one estimation approach for the Gaussian copula model is to reparameterize the model in terms of a non-identifiable covariance matrix Σ, but focus our posterior inference on the identifiable correlation matrix $\Psi = h(\Sigma)$.

This technique of modeling in terms of a non-identifiable parameter in order to simplify calculations is referred to as parameter expansion, and has been used in the context of modeling multivariate ordinal data.

To summarize, we will base our posterior distribution on

$$\begin{aligned} \Sigma &\sim \text{inverse-Wishart}(\nu_0, \mathbf{S}_0^{-1}) \\ \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \Sigma) \\ Y_{i,j} &= g_j(Z_{i,j}), \end{aligned} \tag{12.7}$$

but our estimation and inference will be restricted to $\Psi = h(\Sigma)$. Interestingly, the posterior distribution for Ψ obtained from this prior and model is exactly the same as that which would be obtained from the following:

$$\begin{aligned} \Sigma &\sim \text{inverse-Wishart}(\nu_0, \mathbf{S}_0^{-1}) \\ \Psi &= h(\Sigma) \\ \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n &\sim \text{i.i.d. multivariate normal}(0, \Psi) \\ Y_{i,j} &= g_j(Z_{i,j}). \end{aligned} \tag{12.8}$$

The non-identifiable model (Eq12.7) gives the same posterior distribution for Ψ as the identifiable model in Eq12.8 in which Ψ prior distribution is defined by $\{\Sigma \sim \text{IW}(\nu_0, \mathbf{S}_0^{-1}), \Psi = h(\Sigma)\}$. Only difference: the Gibbs sampling scheme for Eq12.7 is easier to formulate. The equivalence of these models relies on the scale invariance of the rank likelihood, and so will not generally hold for other types of models involving correlation matrices.

Full conditional distribution of Σ

If the prior distribution for Σ: inverse-Wishart$(\nu_0, \mathbf{S}_0^{-1})$, then, as Section 7.3, the full conditional distribution of Σ is also inverse-Wishart. Noting that the probability density of the $n \times p$ matrix $\mathbf{Z}$ can be written as

$$p(\mathbf{Z} \mid \Sigma) = \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2} z_i \Sigma^{-1} z_i\} = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\text{tr}(\mathbf{Z}^T \mathbf{Z} \Sigma^{-1})/2\},$$

The full conditional distribution $p(\Sigma \mid \mathbf{Z}, \mathbf{Z} \in R(\mathbf{Y})) = p(\Sigma \mid \mathbf{Z})$ is then given by

$$\begin{aligned} p(\Sigma \mid \mathbf{Z}) &\propto p(\Sigma) \times p(\mathbf{Z} \mid \Sigma) \\ &\propto |\Sigma|^{-(\nu_0+p+1)/2} \exp\{-\text{tr}(\mathbf{S}_0 \Sigma^{-1})/2\} \times |\Sigma|^{-n/2} \exp\{-\text{tr}(\mathbf{Z}^T \mathbf{Z} \Sigma^{-1})/2\} \\ &= |\Sigma|^{-([\nu_0+n]+p+1)/2} \exp\{-\text{tr}([\mathbf{S}_0 + \mathbf{Z}^T \mathbf{Z}] \Sigma^{-1}/2\} \end{aligned}$$

which is proportional to an inverse-Wishart $(\nu_0 + n, [\mathbf{S}_0 + \mathbf{Z}^T \mathbf{Z}]^{-1})$ density.

Full conditional distribution of $\boldsymbol{Z}$

Section 7.6: If $\boldsymbol{Z}$ is a random multivariate normal$(0, \Sigma)$ vector, then the conditional distribution of Z_j , given the other elements $\boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}$, is a univariate normal distribution with mean and variance given by

$$\begin{aligned} \text{E}[Z_j \mid \Sigma, \boldsymbol{z}_{-j}] &= \Sigma_{j,-j} (\Sigma_{-j,-j})^{-1} \boldsymbol{z}_{-j} \\ \text{Var}[Z_j \mid \Sigma, \boldsymbol{z}_{-j}] &= \Sigma_{j,j} - \Sigma_{j,-j} (\Sigma_{-j,-j})^{-1} \Sigma_{-j,j}, \end{aligned}$$

where $\Sigma_{j,-j}$ refers to the jth row of Σ with the jth column removed, and $\Sigma_{-j,-j}$ refers to Σ with both the jth row and column removed. If in addition we condition on the information that $\mathbf{Z} \in R(\mathbf{Y})$, then we know that $\max\{z_{k,j} : y_{k,j} < y_{i,j}\} < Z_{i,j} < \min\{z_{k,j} : y_{i,j} < y_{k,j}\}$. These two pieces of information imply that the full conditional distribution of $Z_{i,j}$ is a constrained normal distribution.

12.3 Discussion and further references

Normally distributed latent variables are often used to induce dependence among a set of non-normal observed variables.

The rank likelihood is based on the marginal distribution of the ranks, and so is called a marginal likelihood. Marginal likelihoods are typically constructed so that they use the information in the data that depends only on the parameters of interest, and do not use any information that depends on nuisance parameters. Marginal likelihoods do not generally provide efficient estimation, as they throw away part of the information in the data. However, they can turn a very difficult semiparametric estimation problem into essentially a parametric one.