

BIOS 779: Bayesian Statistics

Joseph G. Ibrahim

Department of Biostatistics
University of North Carolina at Chapel Hill

Spring, 2023

Chapter 1:

Introduction to Bayesian Methods

The Bayesian Paradigm

We develop the Bayesian paradigm for parametric inference. To this end, suppose we conduct (or wish to design) a study, in which the parameter θ is of inferential interest. Here θ may be vector valued. For example,

- 1) θ = difference in treatment means
- 2) θ = hazard ratio
- 3) θ = vector of regression coefficients
- 4) θ = probability that the treatment is effective

In parametric inference, we specify a parametric model for the data, indexed by the parameter θ . Letting x denote the data, we denote this model (density) by $p(x | \theta)$. The likelihood function of θ is **any** function proportional to $p(x | \theta)$, i.e.,

$$L(\theta) \propto p(x | \theta) .$$

Example 1.1

Suppose $x \mid \theta \sim \text{Binomial}(N, \theta)$. Then

$$p(x \mid \theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}, \quad x = 0, 1, \dots, N.$$

We can take

$$L(\theta) = \theta^x (1 - \theta)^{N-x}.$$

The parameter θ is unknown. In the Bayesian mind-set, we express our uncertainty about quantities by specifying distributions for them. Thus, we express our uncertainty about θ by specifying a **prior distribution** for it. We denote the prior density of θ by $\pi(\theta)$.

The word “prior” is used to denote that it is the density of θ before the data x is observed. By Bayes’ theorem, we can construct the distribution of $\theta \mid x$, which is called the **posterior distribution** of θ .

We denote the posterior distribution of θ by $p(\theta | x)$.

By Bayes theorem,

$$p(\theta | x) = \frac{p(x | \theta) \pi(\theta)}{\int_{\Theta} p(x | \theta) \pi(\theta) d\theta} , \quad (1.1)$$

where Θ denotes the parameter space of θ . The quantity

$$p(x) = \int_{\Theta} p(x | \theta) \pi(\theta) d\theta$$

is the normalizing constant of the posterior distribution of θ . $p(x)$ is also the **marginal distribution** of x , and is sometimes called the marginal distribution of the data.

For most inference problems, $p(x)$ does not have a closed form. Bayesian inference about θ is primarily based on the posterior distribution of θ , $p(\theta | x)$. For example, one can compute various posterior summaries, such as the mean, median, mode, variance, and quantiles. For example, the posterior mean of θ is given by

$$E(\theta | x) = \int_{\Theta} \theta p(\theta | x) d\theta .$$

The posterior mode of θ is the value of θ that maximizes $p(\theta | x)$.

Example 1.2

Given θ , suppose x_1, \dots, x_n are i.i.d. Binomial($1, \theta$), and $\theta \sim \text{beta}(\alpha, \lambda)$. The parameters of the prior distribution are often called the **hyperparameters**. Let us derive the posterior distribution of θ .

Let $x = (x_1, \dots, x_n)$, and thus

$$\begin{aligned} p(x \mid \theta) &= \prod_{i=1}^n p(x_i \mid \theta) \\ &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}, \end{aligned}$$

where $\sum x_i = \sum_{i=1}^n x_i$. Also,

$$\pi(\theta) = \frac{\Gamma(\alpha + \lambda)}{\Gamma(\alpha)\Gamma(\lambda)} \theta^{\alpha-1} (1-\theta)^{\lambda-1}.$$

Now, we can write the **kernel** of the posterior density as

$$\begin{aligned} p(\theta | x) &\propto \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \theta^{\alpha-1} (1-\theta)^{\lambda-1} \\ &= \theta^{\sum x_i + \alpha - 1} (1-\theta)^{n - \sum x_i + \lambda - 1}. \end{aligned}$$

Thus

$$p(\theta | x) \propto \theta^{\sum x_i + \alpha - 1} (1-\theta)^{n - \sum x_i + \lambda - 1}.$$

We can recognize this kernel as a beta kernel with parameters $(\sum x_i + \alpha, n - \sum x_i + \lambda)$. Thus

$$\theta | x \sim \text{beta}(\sum x_i + \alpha, n - \sum x_i + \lambda),$$

and therefore

$$\begin{aligned} p(\theta | x) &= \frac{\Gamma(\alpha + n + \lambda)}{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \lambda)} \\ &\times \theta^{\sum x_i + \alpha - 1} (1-\theta)^{n - \sum x_i + \lambda - 1}. \end{aligned}$$

Remark 1.1

In deriving posterior densities, an often used technique is to try and recognize the kernel of the posterior density of θ . This avoids a direct computation of $p(x)$. This technique saves lots of time in the derivation. If the kernel cannot be recognized, then $p(x)$ must be computed directly.

In this example, we have

$$\begin{aligned} p(x) &= p(x_1, \dots, x_n) \\ &= \frac{\Gamma(\alpha + \lambda)}{\Gamma(\alpha)\Gamma(\lambda)} \int_0^1 \theta^{\sum x_i + \alpha - 1} (1 - \theta)^{n - \sum x_i + \lambda - 1} d\theta \\ &= \frac{\Gamma(\alpha + \lambda)}{\Gamma(\alpha)\Gamma(\lambda)} \frac{\Gamma(\sum x_i + \alpha)}{\Gamma(n - \sum x_i + \lambda)} \frac{\Gamma(n - \sum x_i + \lambda)}{\Gamma(\alpha + n + \lambda)} . \end{aligned}$$

Thus

$$p(x_1, \dots, x_n) = \frac{\Gamma(\alpha + \lambda)}{\Gamma(\alpha)\Gamma(\lambda)} \frac{\Gamma(\sum x_i + \alpha)}{\Gamma(\alpha + n + \lambda)} \frac{\Gamma(n - \sum x_i + \lambda)}{\Gamma(n - \sum x_i + \lambda)} ,$$

for $x_i = 0, 1, \dots, n$.

Suppose A_1, A_2, \dots are events such that $A_i \cap A_j = \phi$ and $\bigcup_{i=1}^{\infty} A_i = \Omega$, where Ω denotes the **sample space**. Let B denote an event in Ω .

Then Bayes theorem for events can be written as

$$P(A_i | B) = \frac{P(B | A_i) P(A_i)}{\sum_{j=1}^{\infty} P(B | A_j) P(A_j)} .$$

We can compare this formula with that for random variables given in (1.1). $P(A_i)$ is the prior probability of A_i and $P(A_i | B)$ is the posterior probability of A_i given B has occurred.

Example 1.3

Bayes theorem is often used in diagnostic tests for cancer.

A young person was diagnosed as having a type of cancer that occurs extremely rarely in young people. Naturally, he was very upset. A friend told him that it was probably a mistake. His friend reasoned as follows. No medical test is perfect: there are always incidences of false positives and false negatives.

Let C stand for the event that he has cancer and let $+$ stand for the event that an individual responds positively to the test. Assume

$P(C) = 1/1,000,000 = 10^{-6}$, $P(+) | C) = .99$, and $P(+) | C^c) = .01$. (So only one per million people his age have the disease and the test is extremely good relative to most medical tests, giving only 1% false positives and 1% false negatives.)

Find the probability that he has cancer given that he has a positive response. (After you make this calculation you will not be surprised to learn that he did not have cancer.)

$$\begin{aligned} P(C | +) &= \frac{P(+) | C) P(C)}{P(+) | C) P(C) + P(+) | C^c) P(C^c)} \\ &= \frac{(.99)(10^{-6})}{(.99)(10^{-6}) + (.01)(.999999)} \\ &= \frac{.00000099}{.01000098} = \boxed{.00009899} \end{aligned}$$

Example 1.4: Deciding Paternity

Legal cases of disputed paternity in many countries are resolved using blood tests. Laboratories make genetic determinations concerning the mother, child, and alleged father. (Some cases involve different types of evidence. The mother may not be available. The alleged father may not be available, but his brother is available, and so on.)

Most labs apply Bayes rule in communicating the testing results. They calculate the probability that the alleged father is in fact the child's father given the genetic evidence. (You should ask yourself, Whose probability is it?)

For the sake of brevity, we will pare down the genetic evidence usually introduced and deal only with ABO blood type. Knowing some genetics may help you to follow this example, but such knowledge is not required. All the probabilities you need will be given.

Suppose you are on a jury considering a paternity suit brought by Suzy Smith's mother against Al Edged. (Many paternity suits in the United States are initiated by welfare departments in the name of the mother.) The following is part of the background information: Suzy's mother has blood type O and Al Edged is type AB. All your probabilities are calculated conditional on this information, although we will not include it explicitly to the right of the vertical bar in probability expressions.

You have other information as well. You hear testimony concerning whether Al Edged and Suzy's mother had sexual intercourse during the time that conception could have occurred, about the timing and frequency of such intercourse, about Al Edged's fertility, about the possibility that someone else is the father, and so on. You put all this information together in assessing $P(F)$, your probability that Al is Suzy's father.

The evidence of interest is Suzy's blood type. If it is O, then Al Edged is excluded from paternity—he is not her father, unless there has been a gene mutation or a laboratory error. Suzy's blood type turns out to be type B; call this event B . According to Bayes rule,

$$P(F \mid B) = \frac{P(B \mid F)P(F)}{P(B \mid F)P(F) + P(B \mid F^c)P(F^c)} .$$

According to Mendelian genetics, $P(B \mid F) = \frac{1}{2}$. You also accept the blood bank's $P(B \mid F^c)$. They calculate this as the proportion of B genes (this is not the same as the proportion of people with type B blood) to the total number of ABO genes in their previous cases.

A typical value among Caucasians is 9%. so,

$$\begin{aligned} P(F | B) &= \frac{(1/2)P(F)}{(1/2)P(F) + (.09)P(F^c)} = \\ &\frac{50 \cdot P(F)}{41 \cdot P(F) + 9} . \end{aligned}$$

This is a substantial increase over $P(F)$. For example, it is about 85% when $P(F) = \frac{1}{2}$. The reason such a large increase is possible is that Suzy's paternal gene (B) is relatively rare. The probability of paternity would increase for any male who has a B gene.

The relationship between our unconditional probability, $P(F)$, and our conditional probability, $P(F | B)$, can be shown using the following table:

$P(F)$	0	.100	.250	.500	.750	.900	1
$P(F B)$	0	.382	.649	.847	.943	.980	1

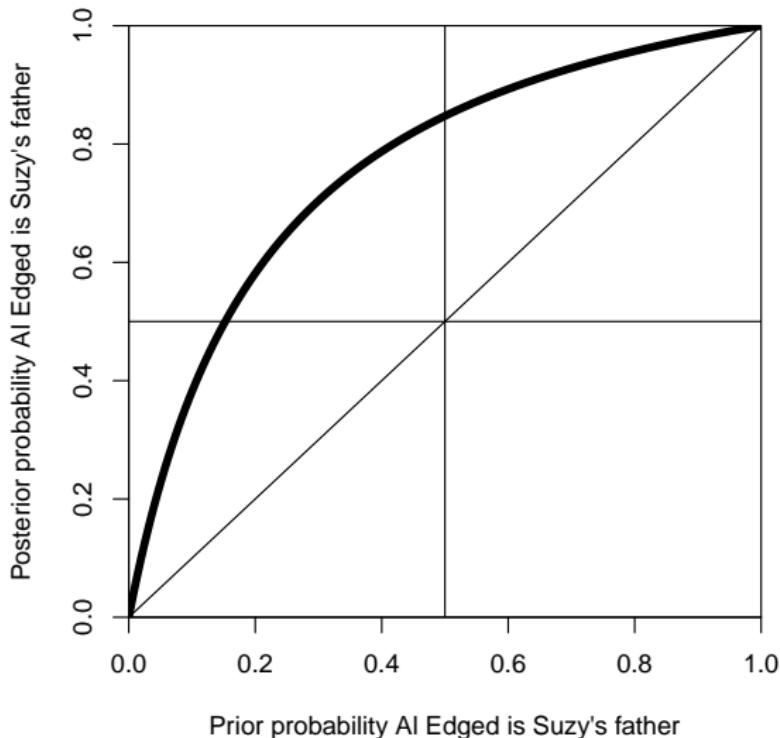


Figure 1.1: Posterior vs. prior probabilities ($P(F|B)$ and $P(F)$) for Example 1.3

Another way to show the same thing is to use a graph, such as the one in Figure 1.1. The diagonal on this graph corresponds to evidence which contains no information about F (in which case B and F would be independent). Comparing this diagonal with the actual curve shows how much the evidence changes one's prior probability of paternity. Tables and graphs are effective ways for juries and others to understand the strength of the evidence.

Blood banks and other laboratories that analyze genetic factors in paternity cases have a name for the Bayes factor in favor of F :

$$\text{Paternity index} = PI = \frac{P(B | F)}{P(B | F^c)} = \frac{1/2}{.09} = 5.56 .$$

The evidence (child has type B blood) is 5.56 times as likely if Al Edged is the father than if he is not.

The posterior probability of paternity (based on the equivalent version of Bayes' rule) is

$$\begin{aligned} P(F | B) &= \frac{1}{1 + \frac{P(B|F^c)}{P(B|F)} \frac{P(F^c)}{P(F)}} \\ &= \frac{1}{1 + \frac{\frac{1}{PI} \frac{P(F^c)}{P(F)}}{}} \\ &= \frac{PI}{PI + \frac{P(F^c)}{P(F)}} . \end{aligned}$$

Laboratories choose $P(F) = \frac{1}{2}$ and report a probability (or likelihood) of paternity as though there is no prior probability involved. This policy is arbitrary and misleading.

For example, on February 18, 1994, news wire services carried this report: “Experts who examined blood samples from John Bobbitt, Beatrice Williams and her son, Andrew, determined there is a 99.99 percent likelihood that Bobbitt is the boy’s father.”

This is stated as though it is a posterior probability $P(F | \text{blood data})$, but it is such only if the other evidence in the case weights equally on F as on F^c from the perspective of the jury. In effect, the experts are preempting the jury in assessing this evidence.

To find the paternity index in the Bobbitt case, carry out the calculations in reverse: the answer is 9,999. Check this, using the formula for probability in terms of PI and assuming $P(F) = .5$ and, therefore, $P(F^c)/P(F) = 1$:

$$P(F | B) = \frac{PI}{1 + PI} = \frac{9,999}{1 + 9,999} = .9999$$

[There is only one significant digit in calculating the PI to be 9,999. Any PI close to 10,000 gives a $P(F | B)$ that rounds to .9999.]

A PI as large as 10,000 is possible only by combining evidence from many genetic factors in addition to ABO blood type. These factors are combined by multiplying their individual likelihoods, which is appropriate if the factors are independent.

Remark 1.2

$\frac{P(B|F)}{P(B|F^c)}$ is called the Bayes Factor in favor of F .

$$\begin{aligned}\text{Bayes factor} &= \text{ratio of posterior odds to prior odds} \\ &= \frac{P(F|B)/P(F)}{P(F^c|B)/P(F^c)} = \frac{P(B|F)}{P(B|F^c)}\end{aligned}$$

If

$1 \leq \text{Bayes factor} \leq 3$	weak evidence in favor of F
$3 \leq \text{Bayes factor} \leq 12$	positive evidence in favor of F
$12 \leq \text{Bayes factor} \leq 150$	strong evidence in favor of F
$\text{Bayes factor} > 150$	decisive evidence in favor of F

Example 1.5

Treating Leukemia

A study reported in 1963 was designed to evaluate the effectiveness of a chemotherapeutic agent, 6-mercaptopurine (6-MP), for the treatment of acute leukemia. Such an evaluation requires a comparison group. Patients were randomized to therapy groups by coin tosses. The first patient was assigned to the 6-MP group if the coin came up heads and to the placebo group if it came up tails; the second patient received the other therapy. This process was repeated for the third and fourth patients, fifth and sixth patients and so on. For each pair of patients, the investigators recorded whether the 6-MP patient or the placebo patient stayed in remission longer.

There were 21 patients in the study. The results were as follows (where B means that the 6-MP patient fared **Better** than the placebo patient in the pair and W means the 6-MP patient fared **Worse**):

BWB BBW BBBB BBBWB BBBB B

So 6-MP was more effective in 18 of the 21 pairs of patients. Thus, the proportion of pairs in which it was better is $18/21 = 85.7\%$.

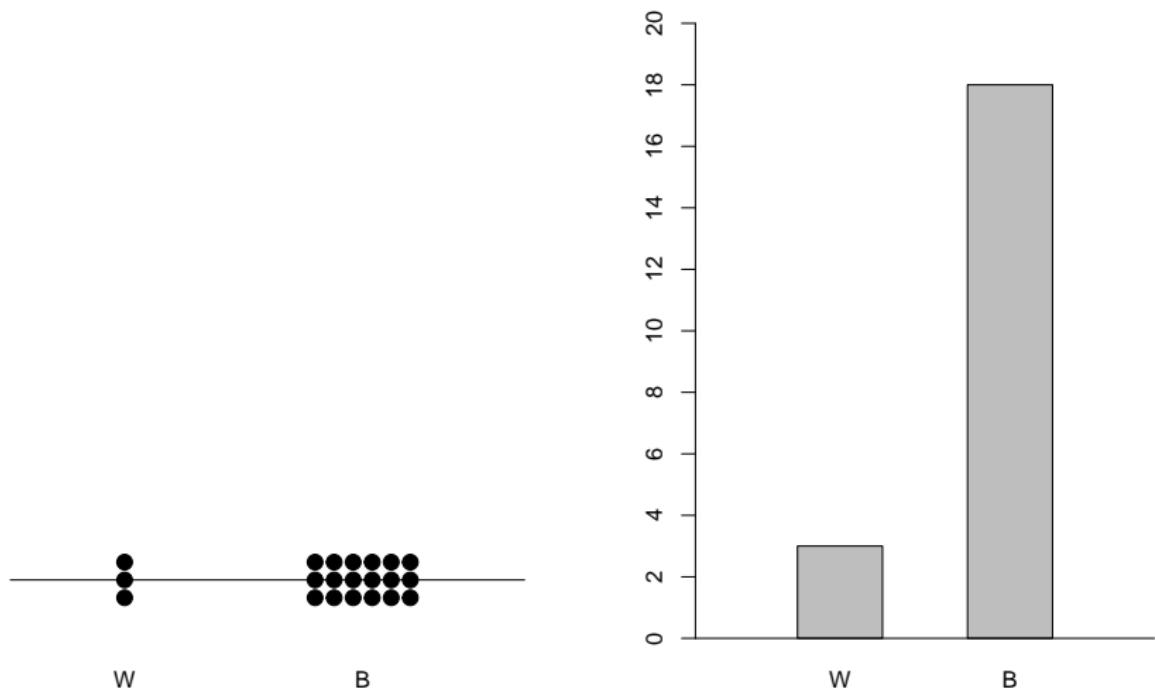


Figure 1.2: Dot plot and bar chart for Example 1.4

Proportions are numbers between 0 and 1. So proportions can be thought of as probabilities. Suppose I label chips with the symbols from the list of B's and W's—18 chips are labeled B and three are label W—and place them in a bowl. I do this in such a way that the chips are exchangeable for me. So when I select a chip from the bowl, my probability of a B is $\frac{18}{21}$. When confined to the 21 pairs in the experiment, my probability that 6-MP is better than placebo is $\frac{18}{21}$.

But the study was wasted if the strongest conclusion one can make applies only to the patients in the study. The important question is whether we can say anything about patients who are candidates for treatment with 6-MP but are not in the study. Consider someone who has acute leukemia and whom we regard as exchangeable with the patients in the trial before they were treated. Suppose that the only possible treatments are 6-MP and placebo (the latter being essentially no treatment). Should this patient be given 6-MP?

This is not an easy question to answer. It involves weighing any possible beneficial effects with possible adverse effects and with other costs (including monetary costs) of 6-MP therapy. Two quite rational patients might select differently. To help with the decision, let's simplify things. Restrict consideration to finding the probability that this next patient will stay in remission longer if treated with 6-MP than if not treated. (Assuming that the patients are exchangeable, this is the same as saying that for the next pair of patients treated, one with 6-MP and the other with placebo, the 6-MP patient will stay in remission longer.)

What is the posterior probability that the 6-MP is effective in prolonging remission?

We need to talk about population models that are analogous to treating leukemia patients. Suppose the population models contain B's and W's. Selecting a B is equivalent to the 6-MP patient being in remission longer. The prior probabilities of the various distributions of B's and W's depend on the person making the assessment. An important person in this regard is a clinician who was involved in the study.

I do not know what the clinicians thought about the effectiveness of 6-MP before the study. They must have had some reason to think that it was effective or they would not have organized the study. They also must have given some prior probability of the possibility that 6-MP was not effective or they could not have condoned giving a placebo to some of the patients.

I am going to specify *two* different prior probability distributions, one for Dr. X and the other for Dr. Y. While it is possible that one of these corresponded to an actual clinician's opinions before the study, I do not know whether it is so. My main purpose is to show how someone's prior opinion is updated using data in the study. A secondary purpose is to show you how much more closely the two physicians agree after they observe the data than they did before.

Suppose the total number of chips in the models is 10, with either 0, 1, 2, up to 10 B's. When things get crowded, I will let p stand for the proportion of B's, which then is either 0, .1, .2, up to 1. If $p > .5$, then 6-MP is effective, and if $p < .5$, then it is detrimental. The null hypothesis of no effect is $p = .5$.

Dr. X is quite open-minded about the effectiveness of 6-MP, whereas Dr. Y is somewhat more optimistic about its effectiveness. Both feel that it is possible that 6-MP is detrimental, but Dr. X rates this as more likely than Dr. Y.

Their probabilities are given in Figure 1.3, tabulated in Table 1.1, and presented in bar chart form in Figure 1.4.

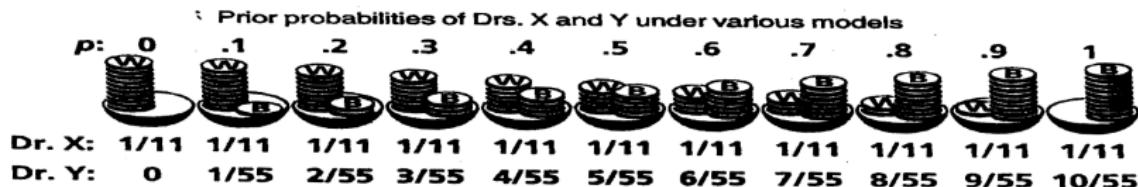


Figure 1.3:

Table 1.1: Prior probabilities of Drs. X and Y

Proportion of B's	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
Dr. X's prior	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11
Dr. Y's prior	0	1/55	2/55	3/55	4/55	5/55	6/55	7/55	8/55	9/55	10/55

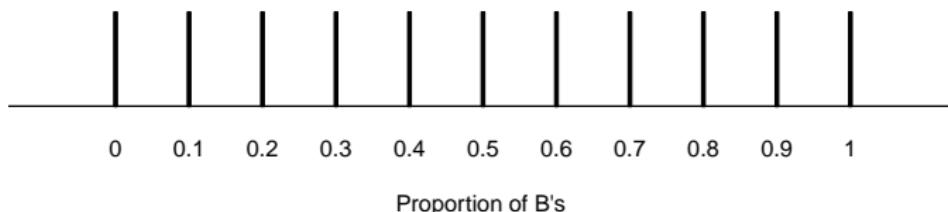
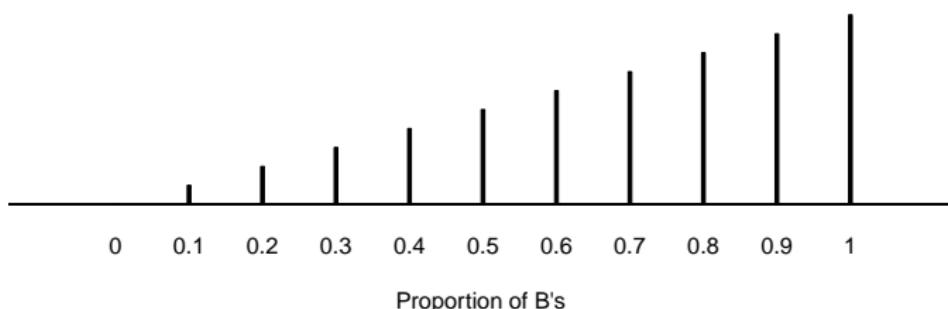
Prior probabilities of Dr. X**Prior probabilities of Dr. Y**

Figure 1.4: Bar charts of prior probabilities of Drs. X and Y

Consider the first pair of patients, one assigned to 6-MP and the other to placebo. By the law of total probability, Dr. X's predictive probability of event B—the 6-MP patient will stay in remission longer—is

$$\begin{aligned} P(\text{B for Dr. X}) &= 0 \times \frac{1}{11} + .1 \times \frac{1}{11} + .2 \times \frac{1}{11} + .3 \times \frac{1}{11} + \cdots + 1 \times \frac{1}{11} \\ &= \frac{1}{2} \end{aligned}$$

For Dr. Y,

$$\begin{aligned} P(\text{B for Dr. Y}) &= 0 \times 0 + .1 \times \frac{1}{55} + .2 \times \frac{2}{55} + .3 \times \frac{3}{55} + \cdots + 1 \times \frac{10}{55} \\ &= \frac{7}{10} \end{aligned}$$

These are the means of the corresponding bar charts (Figure 1.4). The mean for Dr. X is .5, whereas for Dr. Y it is .7. Obviously, they had different prior information, or at least they assessed the available information differently.

What happens when they learn the results of the study? In particular, do they come closer together in their opinions? We want to find the likelihoods. Suppose that Drs. X and Y both regard the pairs in the study to be exchangeable. (This means, for example, that it is as likely that 6-MP will be better in the first pair as in any other pair. However, if the first pair results in B, then the probability that the second results in B may well increase.) The data are as follows

B W B B B W B B B B B B B B B W B B B B B B B B

The likelihood is the probability of the data. The probability of the data depends on the model, that is, on the proportion of B's in the model, which we have called p . To be specific, consider $p = .8$. The likelihood of $p = .8$ is the probability of the data assuming the model has eight B's and two W's. This is the product of the 21 separate probabilities:

(.8)(.2)(.8)(.8)(.8)(.2)(.8)(.8)(.8)(.8)(.8)(.8)(.8)(.8)(.2)(.8)(.8)(.8)(.8)(.8)(.8)(.8)

This equals $(.8)^{18}(.2)^3 = .000144 = 1.44\text{E-}4$. (Exponential notation E-4 means to move the decimal point four places to the left. Similarly, E-13 would mean to move the decimal point 13 places to the left and E+13 would mean to move the decimal point 13 places to the right.)

Likelihoods for all 11 models are shown in the third column of Table 1.2. The table applies for Dr. X and shows the calculations that lead to posterior probabilities—the next to last column in the table. Numbers too small to matter in the end are called “tiny”. My practice is to carry a high degree of accuracy except when writing down a final answer. This guarantees that when I write down the three-digit final answer, it is accurate to three digits. If you check my calculations using only accuracy in the table, you will get slightly different answers. For example, in the row $p = .9$, $1.36\text{E-}5$ divided by $3.14\text{E-}5$ gives .433 instead of .435. That is because $1.36\text{E-}5$ is rounded off from $1.3645\text{E-}5$ and $3.14\text{E-}5$ is rounded off from $3.1379\text{E-}5$.

Table 1.2: Calculations for Dr. X

Model	Prior	Likelihood	Prior × Likelihood	Posterior	Model × Posterior
0	1/11	0	0	0	0
.1	1/11	tiny	tiny	tiny	0
.2	1/11	tiny	tiny	tiny	0
.3	1/11	tiny	tiny	tiny	0
.4	1/11	tiny	tiny	tiny	0
.5	1/11	4.77E-7	4.33E-8	.001	.001
.6	1/11	6.50E-6	5.91E-7	.019	.011
.7	1/11	4.40E-5	4.00E-6	.127	.089
.8	1/11	1.44E-4	1.31E-5	.428	.334
.9	1/11	1.50E-4	1.36E-5	.435	.391
1	1/11	0	0	0	0
Sum	1		3.14E-5	1	.826

In Table 1.2, the sum of the fourth column ($3.14E-5$) is $P_X(D)$. The fifth column contains Dr. X's posterior probabilities, calculated by dividing the entries in the fourth column by the sum of the entries in the fourth column. Because the prior probabilities are equal, the third, fourth, and fifth columns in this table are proportional to each other. So a picture of the likelihoods is the same as the bar chart of Dr. X's posterior probabilities—see Figure 1.5. The last column of the table served to find the mean of this bar chart, which is .826.

Table 1.3 gives the analogous quantities for Dr. Y. The likelihoods are the same as for Dr. X, so the third column of Dr. Y's table is the same. Dr. Y's posterior probabilities (fifth column in the table) are shown in the bar chart in Figure 1.6.

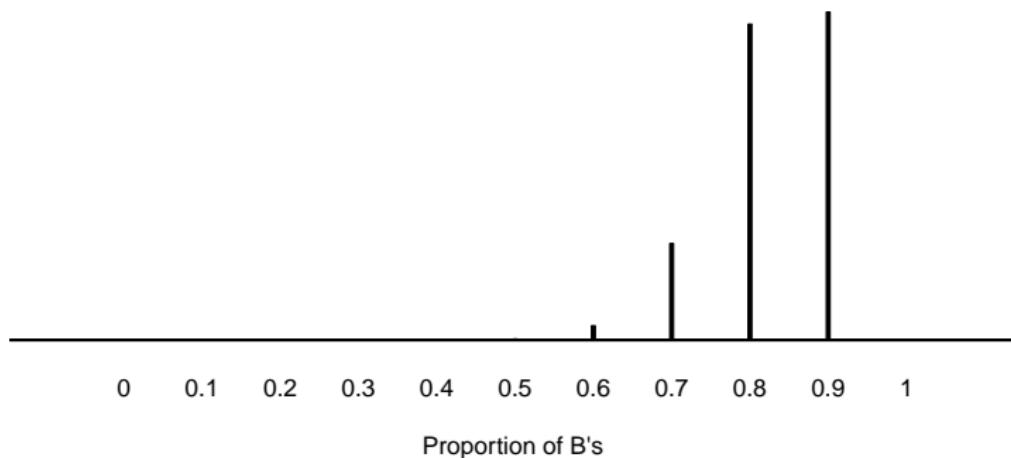


Figure 1.5: Bar chart of Dr. X's posterior probabilities (proportional to likelihoods)

Table 1.3: Calculations for Dr. Y

Model	Prior	Likelihood	Prior × Likelihood	Posterior	Model × Posterior
0	0	0	0	0	0
.1	1/55	tiny	tiny	tiny	0
.2	2/55	tiny	tiny	tiny	0
.3	3/55	tiny	tiny	tiny	0
.4	4/55	tiny	tiny	tiny	0
.5	5/55	4.77E-7	4.33E-8	.001	.001
.6	6/55	6.50E-6	7.09E-7	.014	.008
.7	7/55	4.40E-5	5.60E-6	.108	.076
.8	8/55	1.44E-4	2.10E-5	.405	.324
.9	9/55	1.50E-4	2.45E-5	.472	.425
1	10/55	0	0	0	0
Sum	1		5.19E-5	1	.834

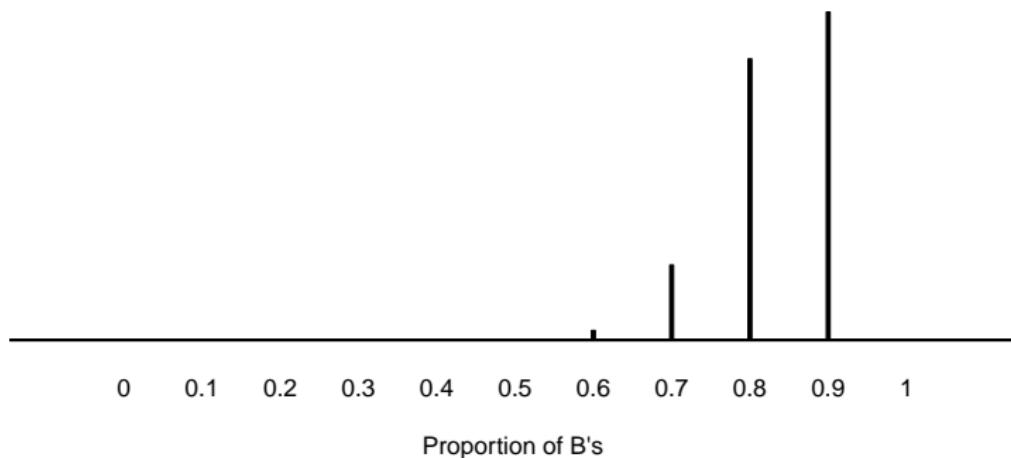


Figure 1.6: Bar chart of Dr. Y's posterior probabilities (proportional to likelihoods)

The posterior bar charts of Drs. X and Y are quite similar, even though their prior bar charts are different. For example, both now associate a posterior probability of .999 with 6-MP having a beneficial effect (this is the sum of the probabilities of proportions of B's greater than .5). Also, the two means (the predictive probabilities that the next chip drawn is a B) are the same to two decimal places:

$$\text{For Dr. X: } P(B|D) = .826$$

$$\text{For Dr. Y: } P(B|D) = .834$$

In terms of the clinical study, these are their probabilities that the next patient will stay in remission longer if treated with 6-MP. The small differences in these two values is due to the difference in the two prior probability distributions.

So even though Drs. X and Y started out far apart (with means of .5 and .7), their posterior probabilities are quite similar. They changed their views as they observed the data. This is possible because both were reasonably open-minded about the effectiveness of the treatment. In terms of population models, they allowed for various possible proportions, p .

Consider a know-it-all—call him Dr. Z—is singularly convinced of the null hypothesis that 6-MP has no effect. Dr. Z “knows” that the better response within each pair of patients is as likely to be on placebo as on 6-MP. Dr. Z’s prior probabilities would not change for *any* data—which is what I mean by know-it-all.

Thus, Dr. Z has a single type of population model in mind, one with proportion p equal to 0.5. One such model has five chips of each type.



Figure 1.7:

Table 1.4: Calculations for Dr. Z

Model	Prior	Likelihood	Prior × Likelihood	Posterior	Model × Posterior
.5	1	4.77E-7	4.77E-7	1	.500
Sum	1		4.77E-7	1	.500

The probability table of Dr. Z (Table 1.4) is analogous to those for Dr. X and Y. Calculations using Bayes' rule are now trivial.

To be thoroughly convinced as Dr. Z would require impossibly strong prior evidence. Only a fool is certain. One's probability may reasonably be very large or very small, but they should not be 1 or 0. I seem to be ignoring my own advice by giving probability 1 to sets of models. Learning can take place only in the restricted framework of a model universe. We *condition* on the correct model being in a particular universe—and learning is relative to this universe. But we can never *know* that our universe contains the true model. Considering a larger universe may give greater confidence that the model is contained therein, but there is no all-encompassing universe containing all possible models.

Being called a fool convinced Dr. Z to back off from his prior certainty of the null hypothesis. He still feels strongly, but he has backed off from being 100% sure to 99% sure. He spreads his remaining 1% evenly among the other 10 possible p 's. The initial table of probabilities changed accordingly—see Table 1.5. While 100% certainty was not changed by the data, the 99% probability of the null hypothesis drops to about 58%. Dr. Z's probabilities that the next patient will stay in remission longer if treated with 6-MP has increased from 50% (no benefit) to about 64%.

While Drs. X and Y were somewhat different in their prior views, they were quite far from the new Dr. Z. All three came closer together in their views after seeing the data, but Dr. Z is still rather different from the other two. In particular, he still gives a substantial probability (58%) to the null hypothesis. They might resolve their differences by discussing their prior information and the bases for their disagreements. But there is no reason that they should agree.

Table 1.5: Calculations for new Dr. Z

Model	Prior	Likelihood	Prior × Likelihood	Posterior	Model × Posterior
0	.001	0	0	0	0
.1	.001	tiny	tiny	tiny	0
.2	.001	tiny	tiny	tiny	0
.3	.001	tiny	tiny	tiny	0
.4	.001	tiny	tiny	tiny	0
.5	.990	4.77E-7	4.72E-7	.578	.289
.6	.001	6.50E-6	6.50E-9	.008	.005
.7	.001	4.40E-5	4.40E-8	.054	.038
.8	.001	1.44E-4	1.44E-7	.176	.141
.9	.001	1.50E-4	1.50E-7	.184	.166
1	.001	0	0	0	0
Sum	1		8.17E-7	1	.639

What about people other than Drs. X, Y, and Z? How should they decide whom to believe. Consider a regulatory agency (such as the U.S. Food and drug administration) deciding whether to approve 6-MP for standard use based on this study. I cannot do justice to this important issue here, but I will mention some considerations. If the drug is not approved, then more studies would have to be conducted. This would delay delivering possible efficacious therapy to patients not in the studies. If it is approved, then the opportunity for further experimentation may be lost. The benefits suggested by this study may not be real, and an ineffective therapy might be given to thousands of patients who could perhaps be treated better with other therapies.

A consideration is that Dr. Z's attitude may be typical among oncologists and they would not use the drug even if it were approved. Perhaps another study would convince them. Finally, I have considered only effectiveness. Decision to use the drug or not and to approve the drug or not depend on its side effects. A drug that prolongs life might make that life of sufficient low quality that the benefits are more than offset by the drawbacks.

Example 1.6

Suppose x_1, \dots, x_n is a random sample from $N(\mu, \sigma^2)$.

- (i) Suppose σ^2 is known and $\mu \sim N(\mu_0, \sigma_0^2)$. The posterior density of μ is given by:

$$\begin{aligned}
 p(\mu | \sigma^2, x) &\propto \left(\prod_{i=1}^n p(x_i | \mu, \sigma^2) \right) \pi(\mu) \\
 &\propto \left(\exp \left\{ -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right\} \right) \left(\exp \left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \right) \\
 &\propto \exp \left\{ -\frac{1}{2} \left(\frac{n\sigma_0^2 + \sigma^2}{\sigma_0^2 \sigma^2} \right) \mu^2 + 2\mu \left(\frac{\sigma_0^2 \sum x_i + \mu_0 \sigma^2}{2\sigma^2 \sigma_0^2} \right) \right\} \\
 &= \exp \left\{ -\frac{1}{2} \left(\frac{n\sigma_0^2 + \sigma^2}{\sigma_0^2 \sigma^2} \right) \left[\mu^2 - 2\mu \left(\frac{\sigma_0^2 \sum x_i + \mu_0 \sigma^2}{n\sigma_0^2 + \sigma^2} \right) \right] \right\} \\
 &\propto \exp \left\{ -\frac{1}{2} \left(\frac{n\sigma_0^2 + \sigma^2}{\sigma_0^2 \sigma^2} \right) \left[\mu - \left(\frac{\sigma_0^2 \sum x_i + \mu_0 \sigma^2}{n\sigma_0^2 + \sigma^2} \right) \right]^2 \right\}.
 \end{aligned}$$

We can recognize this as a normal kernel with mean $\mu_{\text{post}} = \frac{\sigma_0^2 \sum x_i + \mu_0 \sigma^2}{n\sigma_0^2 + \sigma^2}$ and variance $\sigma_{\text{post}}^2 = \left(\frac{n\sigma_0^2 + \sigma^2}{\sigma_0^2 \sigma^2} \right)^{-1} = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$.

Thus

$$\begin{aligned}\mu|x, \sigma^2 &\sim N\left(\frac{\sigma_0^2 \sum x_i + \mu_0 \sigma^2}{n\sigma_0^2 + \sigma^2}, \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}\right) \\ \mu_{\text{post}} &= \mu_0 \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right) + \bar{x} \left(1 - \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right)\end{aligned}$$

The posterior mean is a weighted average of the sample mean and the prior mean.

Remark 1.3

1. as $n \rightarrow \infty$, $\mu_{\text{post}} \rightarrow \bar{x}$.
2. as $\sigma^2 \rightarrow \infty$, $\mu_{\text{post}} \rightarrow \mu_0$.
3. as $\sigma_0^2 \rightarrow \infty$, $\mu_{\text{post}} \rightarrow \bar{x}$.

- (ii) Suppose μ is known and σ^2 is unknown. Let $\tau = 1/\sigma^2$. τ is often called the **precision** parameter.

Suppose $\tau \sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)$. Thus

$$\pi(\tau) \propto \tau^{\frac{\delta_0}{2}-1} \exp\left\{-\frac{\tau\gamma_0}{2}\right\} .$$

Let us derive the posterior distribution of τ .

$$\begin{aligned} p(\tau | \mu, x) &\propto \left(\tau^{n/2} \exp\left\{-\frac{\tau}{2} \Sigma(x_i - \mu)^2\right\} \right) \left(\tau^{\frac{\delta_0}{2}-1} \exp\left\{-\frac{\tau\gamma_0}{2}\right\} \right) \\ &= \tau^{\frac{n+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2} (\gamma_0 + \Sigma(x_i - \mu)^2)\right\} \end{aligned}$$

Thus

$$\tau | \mu, x \sim \text{gamma}\left(\frac{n + \delta_0}{2}, \frac{\gamma_0 + \Sigma(x_i - \mu)^2}{2}\right) .$$

- (iii) Now suppose μ and σ^2 are both unknown. Suppose we specify the joint prior

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

where $\mu | \tau \sim N(\mu_0, \tau^{-1}\sigma_0^2)$
 $\tau \sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right).$

The joint posterior density of (μ, τ) is given by

$$\begin{aligned} p(\mu, \tau | x) &\propto \left(\tau^{n/2} \exp\left\{-\frac{\tau}{2} \Sigma(x_i - \mu)^2\right\} \right) \left(\tau^{\frac{1}{2}} \exp\left\{-\frac{\tau}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \right) \\ &\quad \times \left(\tau^{\delta_0/2-1} \exp\left\{-\frac{\tau\gamma_0}{2}\right\} \right) \\ &= \tau^{\frac{n+\delta_0+1}{2}-1} \exp\left\{-\frac{\tau}{2} \left(\gamma_0 + \frac{(\mu - \mu_0)^2}{\sigma_0^2} + \Sigma(x_i - \mu)^2 \right) \right\}. \end{aligned}$$

The joint posterior density does not have a clear recognizable form. Thus, we need to compute $p(x)$ by brute force.

$$\begin{aligned}
 p(x) &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+\delta_0+1}{2}-1} \exp\left\{-\frac{\tau}{2} \left(\gamma_0 + \frac{(\mu - \mu_0)^2}{\sigma_0^2} + \Sigma(x_i - \mu)^2\right)\right\} d\mu d\tau \\
 &= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+\delta_0+1}{2}-1} \\
 &\quad \times \exp\left\{-\frac{\tau}{2} \left(\gamma_0 + \mu^2(n + \frac{1}{\sigma_0^2}) - 2\mu(\Sigma x_i + \frac{\mu_0}{\sigma_0^2}) + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2\right)\right\} d\mu d\tau \\
 &= \left(\int_0^\infty \tau^{\frac{n+\delta_0+1}{2}-1} \exp\left\{-\frac{\tau}{2} \left(\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2\right)\right\} \right. \\
 &\quad \times \left. \left(\int_{-\infty}^\infty \exp\left\{-\frac{\tau}{2} \left(\mu^2(n + \frac{1}{\sigma_0^2}) - 2\mu(\Sigma x_i + \frac{\mu_0}{\sigma_0^2})\right)\right\} d\mu \right) d\tau \right).
 \end{aligned}$$

The integral with respect to μ can be evaluated by completing the square.

$$\begin{aligned}
 & \int_{-\infty}^{\infty} \exp \left\{ -\frac{\tau(n + \sigma_0^{-2})}{2} \left[\mu - \frac{(\Sigma x_i + \mu_0 \sigma_0^{-2})}{(n + \sigma_0^{-2})} \right]^2 \right\} \\
 & \quad \times \exp \left\{ \frac{\tau \left[(\Sigma x_i + \mu_0 \sigma_0^{-2})^2 \right]}{2(n + \sigma_0^{-2})} \right\} d\mu \\
 = & \quad \exp \left\{ \frac{\tau \left[(\Sigma x_i + \mu_0 \sigma_0^{-2})^2 \right]}{2(n + \sigma_0^{-2})} \right\} (2\pi)^{1/2} \tau^{-1/2} (n + \sigma_0^{-2})^{-1/2}.
 \end{aligned}$$

Now we need to evaluate

$$\begin{aligned}
 & \int_0^{\infty} (2\pi)^{\frac{1}{2}} (n + \sigma_0^{-2})^{-\frac{1}{2}} \tau^{\frac{n+\delta_0+1}{2}-1} \tau^{-\frac{1}{2}} \\
 & \quad \times \exp \left\{ -\frac{\tau}{2} \left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2 \right] \right\} \exp \left\{ \frac{\tau}{2} \left[\frac{(\Sigma x_i + \frac{\mu_0}{\sigma_0^2})^2}{(n + 1/\sigma_0^2)} \right] \right\} d\tau
 \end{aligned}$$

$$\begin{aligned}
&= (2\pi)^{\frac{1}{2}} \left(n + \frac{1}{\sigma_0^2}\right)^{-\frac{1}{2}} \int_0^\infty \tau^{\frac{n+\delta_0}{2}-1} \\
&\quad \times \exp \left\{ -\frac{\tau}{2} \left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2 - \frac{(\sum x_i + \frac{\mu_0}{\sigma_0^2})^2}{(n + \frac{1}{\sigma_0^2})} \right] \right\} d\tau \\
&= \frac{(2\pi)^{\frac{1}{2}} \Gamma \left(\frac{n+\delta_0}{2}\right) \left(n + \frac{1}{\sigma_0^2}\right)^{-\frac{1}{2}}}{\left[\frac{1}{2} \left(\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2 - \frac{\left(\sum x_i + \frac{\mu_0}{\sigma_0^2}\right)^2}{(n+1/\sigma_0^2)} \right) \right]^{\frac{n+\delta_0}{2}}} \\
&= \frac{(2\pi)^{\frac{1}{2}} \Gamma \left(\frac{n+\delta_0}{2}\right) 2^{\frac{n+\delta_0}{2}} \left(n + \frac{1}{\sigma_0^2}\right)^{-\frac{1}{2}}}{\left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2 - \frac{\left(\sum x_i + \frac{\mu_0}{\sigma_0^2}\right)^2}{(n+1/\sigma_0^2)} \right]^{\frac{n+\delta_0}{2}}} \\
&\propto p(x).
\end{aligned}$$

The joint posterior density of $(\mu, \tau|x)$ can also be obtained in this case by recognizing

$$\begin{aligned} p(\mu, \tau|x) &= p(\mu|\tau, x)p(\tau|x) \\ &= \text{normal} \times \text{gamma} \end{aligned}$$

Exercise 1.1

Find $p(\mu|\tau, x)$, $p(\tau|x)$, and $p(x)$.

It is of greater interest to find the marginal posterior distributions of μ and τ .

$$\begin{aligned}
 p(\mu | x) &= \int_0^\infty p(\mu, \tau | x) d\tau \\
 &\propto \int_0^\infty \tau^{\frac{n+\delta_0+1}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2 \right] \right\} \\
 &\quad \times \exp \left\{ -\frac{\tau}{2} [\mu^2 (n + \frac{1}{\sigma_0^2}) - 2\mu (\Sigma x_i + \frac{\mu_0}{\sigma_0^2})] \right\} d\tau \\
 &= \int_0^\infty \tau^{\frac{n+\delta_0+1}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2 \right] \right\} \\
 &\quad \times \exp \left\{ -\frac{\tau(n + 1/\sigma_0^2)}{2} \left[\left(\mu - \frac{(\Sigma x_i + \mu_0/\sigma_0^2)}{(n + 1/\sigma_0^2)} \right)^2 \right] \right\} \\
 &\quad \times \exp \left\{ \frac{\tau}{2} \left[\frac{(\Sigma x_i + \mu_0/\sigma_0^2)^2}{(n + 1/\sigma_0^2)} \right] \right\} d\tau .
 \end{aligned}$$

Let $a = \frac{(\Sigma x_i + \mu_0/\sigma_0^2)}{(n + 1/\sigma_0^2)}$.

Then, we can rewrite the integral as

$$\begin{aligned}
 & \int_0^\infty \tau^{\frac{n+\delta_0+1}{2}-1} \\
 & \quad \times \exp \left\{ -\frac{\tau}{2} \left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2 + \left(n + \frac{1}{\sigma_0^2} \right) (\mu - a)^2 - \left(n + \frac{1}{\sigma_0^2} \right) a^2 \right] \right\} d\tau \\
 = & \quad \frac{\Gamma \left(\frac{n+\delta_0+1}{2} \right) 2^{\frac{n+\delta_0+1}{2}}}{\left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2 + \left(n + \frac{1}{\sigma_0^2} \right) (\mu - a)^2 - \left(n + \frac{1}{\sigma_0^2} \right) a^2 \right]^{\frac{n+\delta_0+1}{2}}} \\
 \propto & \quad \left[1 + \frac{c(\mu - a)^2}{b - ca^2} \right]^{-\frac{(n+\delta_0+1)}{2}},
 \end{aligned}$$

where $c = n + \frac{1}{\sigma_0^2}$ and $b = \gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \Sigma x_i^2$.

We recognize this kernel as that of a **t distribution** with location parameter a and dispersion parameter $\left(\frac{(n + \delta_0)c}{b - ca^2}\right)^{-1}$, and $n + \delta_0$ degrees of freedom.

Definition 1.1

Let $y = (y_1, \dots, y_p)'$ be a $p \times 1$ random vector. Then y is said to have a p dimensional multivariate t distribution with d degrees of freedom, location parameter m and dispersion matrix $\Sigma_{p \times p}$ if y has density

$$\begin{aligned} p(y) &= \left(\frac{\Gamma(\frac{d+p}{2})(\pi d)^{-p/2} |\Sigma|^{-\frac{1}{2}}}{\Gamma(\frac{d}{2})} \right) \\ &\quad \times \left[1 + \frac{1}{d}(y - m)' \Sigma^{-1} (y - m) \right]^{-\frac{(d+p)}{2}}. \end{aligned}$$

We write this as $y \sim S_p(d, m, \Sigma)$. In our problem, $p = 1$, $d = n + \delta_0$, $m = a$, $\Sigma^{-1} = \frac{(n + \delta_0)c}{b - ca^2}$, $\Sigma = \left(\frac{(n + \delta_0)c}{b - ca^2}\right)^{-1}$.

The marginal distribution of τ is given by

$$\begin{aligned}
 p(\tau | x) &\propto \int_{-\infty}^{\infty} \tau^{\frac{n+\delta_0+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2\right]\right\} \\
 &\quad \times \exp\left\{\frac{\tau}{2}(n + \frac{1}{\sigma_0^2})a^2\right\} \exp\left\{-\frac{\tau(n + \frac{1}{\sigma_0^2})}{2}(\mu - a)^2\right\} d\mu \\
 &\propto \tau^{\frac{n+\delta_0+1}{2}-1} \tau^{-\frac{1}{2}} \exp\left\{-\frac{\tau}{2}\left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2 - (n + \frac{1}{\sigma_0^2})a^2\right]\right\} \\
 &= \tau^{\frac{n+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}\left[\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2 - (n + \frac{1}{\sigma_0^2})a^2\right]\right\}.
 \end{aligned}$$

Thus

$$\tau | x \sim \text{gamma}\left[\frac{n + \delta_0}{2}, \frac{1}{2}\left(\gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum x_i^2 - (n + \frac{1}{\sigma_0^2})a^2\right)\right].$$

Remark 1.4

A t distribution can be obtained as a scale mixture of a normal. That is, if $x \mid \tau \sim N_p(m, \tau^{-1}\Sigma)$ and $\tau \sim \text{gamma}(\frac{\delta_0}{2}, \frac{\gamma_0}{2})$, then

$$p(x) = \int_0^\infty p(x \mid \tau) \pi(\tau) d\tau$$

is the $S_p(\delta_0, m, \frac{\gamma_0}{\delta_0}\Sigma)$ density. That is,

$$x \sim S_p \left(\delta_0, m, \frac{\gamma_0}{\delta_0} \Sigma \right) .$$

Note:

$$\begin{aligned} p(x \mid \tau) &= (2\pi)^{-\frac{p}{2}} \tau^{p/2} |\Sigma|^{-1/2} \\ &\times \exp \left\{ -\frac{\tau}{2} (x - m)' \Sigma^{-1} (x - m) \right\} . \end{aligned}$$

From the earlier derivation, we have

$$\begin{aligned} p(\mu|x) &\propto \left[1 + \frac{c(\mu-a)^2}{b-ca^2}\right]^{-(n+\delta_0+1)/2} \\ &= \left[1 + \frac{(\mu-a)c(\mu-a)}{b-ca^2}\right]^{-(n+\delta_0+1)/2} \end{aligned}$$

where

$$\begin{aligned} c &= n + \sigma_0^{-2} \\ a &= \frac{\sum_{i=1}^n x_i + \mu_0 \sigma_0^{-2}}{n + \sigma_0^{-2}} \\ b &= \gamma_0 + \frac{\mu_0^2}{\sigma_0^2} + \sum_{i=1}^n x_i^2. \end{aligned}$$

We can write nice forms for a and $b - ca^2$ as follows:

$$\begin{aligned} a = \frac{\sum_{i=1}^n x_i + \frac{\mu_0}{\sigma_0^2}}{n + \sigma_0^{-2}} &= \mu_0 \left(\frac{\sigma_0^{-2}}{n + \sigma_0^{-2}} \right) + \bar{x} \left(\frac{n}{n + \sigma_0^{-2}} \right) \\ &= \mu_0 \left(\frac{1}{n\sigma_0^2 + 1} \right) + \bar{x} \left(\frac{n\sigma_0^2}{n\sigma_0^2 + 1} \right) \\ &= \mu_0 w + \bar{x}(1 - w), \quad w = \frac{1}{n\sigma_0^2 + 1}. \end{aligned}$$

Note that $c = n + \sigma_0^{-2} = \sigma_0^{-2}w^{-1}$. Now

$$\begin{aligned}
 b - ca^2 &= \gamma_0 + \mu_0^2\sigma_0^{-2} + \sum_{i=1}^n x_i^2 - \sigma_0^{-2}w^{-1} [\mu_0^2w^2 + \bar{x}^2(1-w)^2 + 2\mu_0\bar{x}w(1-w)] \\
 &= \gamma_0 + \mu_0^2\sigma_0^{-2} + \sum_{i=1}^n x_i^2 - \sigma_0^{-2}\mu_0^2w - \bar{x}^2\sigma_0^{-2}w^{-1}(1-w)^2 - 2\sigma_0^{-2}\mu_0\bar{x}(1-w) \\
 &= \gamma_0 + \mu_0^2\sigma_0^{-2} - \sigma_0^{-2}\mu_0^2w + \sum_{i=1}^n x_i^2 - \bar{x}^2\sigma_0^{-2}w^{-1}(1-w)(1-w) - 2\sigma_0^{-2}\mu_0\bar{x}(1-w) \\
 &= \gamma_0 + \mu_0^2\sigma_0^{-2}(1-w) + \sum_{i=1}^n x_i^2 - \sigma_0^{-2}(1-w) [\bar{x}^2w^{-1}(1-w)] - 2\sigma_0^{-2}\mu_0\bar{x}(1-w) \\
 &= \gamma_0 + \sum_{i=1}^n x_i^2 + \sigma_0^{-2}(1-w) [\mu_0^2 - 2\mu_0\bar{x} + \bar{x}^2] - (1-w)\sigma_0^{-2}\bar{x}^2w^{-1}.
 \end{aligned}$$

Now

$$\begin{aligned}
 \sigma_0^{-2}(1-w) &= \left(1 - \frac{1}{n\sigma_0^2 + 1}\right) \sigma_0^{-2} \\
 &= \left(\frac{n\sigma_0^2 + 1 - 1}{n\sigma_0^2 + 1}\right) \sigma_0^{-2} \\
 &= \left(\frac{n\sigma_0^2 \sigma_0^{-2}}{n\sigma_0^2 + 1}\right) = \frac{n}{n\sigma_0^2 + 1} = nw.
 \end{aligned}$$

Thus

$$\begin{aligned}
 b - ca^2 &= \gamma_0 + \sum_{i=1}^n x_i^2 + nw[\mu_0 - \bar{x}]^2 - nw w^{-1} \bar{x}^2 \\
 &= \gamma_0 + \sum_{i=1}^n x_i^2 - n\bar{x}^2 + nw[\mu_0 - \bar{x}]^2 \\
 &= \gamma_0 + \sum_{i=1}^n (x_i - \bar{x})^2 + nw[\bar{x} - \mu_0]^2 = \tilde{s}^2(n + \delta_0).
 \end{aligned}$$

where $\tilde{s}^2 = (n + \delta_0)^{-1} [\gamma_0 + \sum_{i=1}^n (x_i - \bar{x})^2 + nw(\mu_0 - \bar{x})^2]$.

Thus

$$\mu|x \sim t(n + \delta_0, w\mu_0 + (1 - w)\bar{x}, \tilde{s}^2(n + \sigma_0^{-2})^{-1})$$

where $w = \frac{1}{n\sigma_0^2 + 1}$ and $\tilde{s}^2 = (n + \delta_0)^{-1} [\gamma_0 + \sum_{i=1}^n (x_i - \bar{x})^2 + nw(\bar{x} - \mu_0)^2]$.

Thus

$$\begin{aligned} p(\mu|x) &= \frac{\Gamma\left(\frac{n+\delta_0+1}{2}\right) [\pi(n + \sigma_0^{-2})]^{-1/2} [\tilde{s}^2(n + \sigma_0^{-2})^{-1}]^{-1/2}}{\Gamma\left(\frac{n+\delta_0}{2}\right)} \\ &\quad \times \left[1 + \frac{1}{n + \delta_0} (\mu - a) [\tilde{s}^2(n + \sigma_0^{-2})^{-1}]^{-1} (\mu - a) \right]^{-(n+\delta_0+1)/2}, \end{aligned}$$

where $a = w\mu_0 + (1 - w)\bar{x}$.

Note that

$$\begin{aligned}
 & \frac{1}{n + \delta_0} (\mu - a) [\tilde{s}^2(n + \sigma_0^{-2})^{-1}]^{-1} (\mu - a) \\
 &= \frac{1}{n + \delta_0} \tilde{s}^{-2} (n + \sigma_0^{-2}) (\mu - a)^2 \\
 &= \frac{1}{n + \delta_0} (n + \delta_0) (b - ca^2)^{-1} (n + \sigma_0^{-2}) (\mu - a)^2 \\
 &= (b - ca^2)^{-1} c (\mu - a)^2 \\
 &= \frac{c(\mu - a)^2}{b - ca^2}.
 \end{aligned}$$

Thus,

$$p(\mu|x) \propto \left[1 + \frac{c(\mu - a)^2}{b - ca^2} \right]^{-(n+\delta_0+1)/2}$$

as given earlier.

Remark 1.5

Note that in Examples 1.2 and 1.6 (i), (ii), the posterior distribution is of the same family as the prior distribution. When the posterior distribution of a parameter is of the same family as the prior distribution, such prior distributions are called **conjugate prior distributions**.

For Example 1.2, a beta prior on θ led to a beta posterior for θ . In Example 1.6 i), a normal prior on μ yielded a normal posterior for μ . In Example 1.6 ii), a gamma prior for τ yielded a gamma posterior for τ . We discuss more on conjugate priors later.

Distribution Theory

Theorem 1.1

If $X \sim S_p(v, \mu, \Sigma)$, then $(x - \mu)' \Sigma^{-1} (x - \mu) \sim pF(p, v)$.

Proof:

Recall that if

$$X | \tau \sim N_p(\mu, \tau^{-1} \Sigma)$$

and

$$\tau \sim \text{gamma}(v/2, v/2),$$

then

$$x \sim S_p(v, \mu, \Sigma).$$

Let's find the distribution of

$$(x - \mu)' \Sigma^{-1} (x - \mu) | \tau.$$

Let

$$\begin{aligned}
 z &= \Sigma^{-1/2}(x - \mu), \\
 z|\tau &= N_p(0, \tau^{-1}I), \\
 E(z|\tau) &= \Sigma^{-1/2}E(x - \mu) = 0, \\
 \text{Cov}(z|\tau) &= \Sigma^{-1/2}\text{Cov}(x)\Sigma^{-1/2} \\
 &= \tau^{-1}\Sigma^{-1/2}\Sigma\Sigma^{-1/2} \\
 &= \tau^{-1}I.
 \end{aligned}$$

Now

$$\begin{aligned}
 z'z &= (x - \mu)' \Sigma^{-1} (x - \mu) \\
 &= \sum_{i=1}^p z_i^2 \sim \tau^{-1} \chi_p^2.
 \end{aligned}$$

Since the z_i are i.i.d. $N(0, \tau^{-1})$, we have $z_i^2 \sim \tau^{-1} \chi_1^2$.

Let $R = z'z$, $R|\tau \sim \tau^{-1} \chi_p^2$.

We want to find $f(R)$.

$$\begin{aligned} f(R) &= \int_0^\infty f(R|\tau)f(\tau) d\tau. \\ f(R|\tau) &= \tau f_{\chi_p^2}(R\tau) \\ &\propto \tau(R\tau)^{p/2-1} \exp\left\{-\frac{R\tau}{2}\right\}. \end{aligned}$$

Note

$$\begin{aligned} R &= \tau^{-1}Y, \quad Y \sim \chi_p^2. \\ P(R \leq r|\tau) &= P(\tau^{-1}Y \leq r) \\ &= P(Y \leq \tau r), \\ f(R|\tau) &= \tau f_y(\tau r). \end{aligned}$$

Thus

$$\begin{aligned}
 f(R) &\propto \int_0^\infty \tau (R\tau)^{p/2-1} \exp\left\{-\frac{R\tau}{2}\right\} \tau^{\frac{v}{2}-1} \exp\left\{-\frac{v\tau}{2}\right\} d\tau \\
 &= R^{p/2-1} \int_0^\infty \tau^{\frac{p+v}{2}-1} \exp\left\{-\frac{v\tau}{2}\left(1+\frac{1}{v}R\right)\right\} d\tau \\
 &= R^{p/2-1} \frac{\Gamma(\frac{p+v}{2})}{\left(\frac{v(1+\frac{1}{v}R)}{2}\right)^{(p+v)/2}} \\
 &\propto R^{p/2-1} \left(1 + \frac{1}{v}R\right)^{-(v+p)/2}.
 \end{aligned}$$

Thus

$$f(R) \propto R^{p/2-1} \left(1 + \frac{1}{v}R\right)^{-(v+p)/2}. \quad (1.2)$$

If $S \sim F(a, b)$, then

$$f(s) = \frac{(a/b)^{a/2}}{B(a/2, b/2)} s^{a/2-1} \left(1 + \frac{a}{b}s\right)^{-(a+b)/2}.$$

Equation (1.2) above looks almost like the kernel of an F density. In fact, $R = pR^*$, where $R^* \sim F(p, v)$.

$$\begin{aligned} R^* &= \frac{1}{p}R, \\ f(R^*) &\propto (R^*)^{p/2-1} \left(1 + \frac{p}{v}R^*\right)^{-(v+p)/2} \propto F(p, v). \end{aligned}$$

This proves the theorem.

Theorem 1.2

If $X \sim S_p(v, \mu, \Sigma)$, then as $v \rightarrow \infty$,

$$S_p(v, \mu, \Sigma) \rightarrow N_p \left(\mu, \left(\frac{v}{v-2} \right) \Sigma \right).$$

Advantages of Bayesian Methods

1) Interpretation

Having a distribution for your unknown θ is easier to understand than a point estimate and a standard error. In addition, we consider the following example of a confidence interval. A 95% confidence interval for a population mean θ can be written as

$$\bar{x} \pm (1.96) \frac{s}{\sqrt{n}} .$$

After the sample is collected, either the interval contains θ or it doesn't. Thus

$$P(a < \theta < b) \neq .95 .$$

We have to rely on a repeated sampling interpretation to make a probability statement as above.

Thus, after observing the data, we **cannot** make a statement like “the true θ has a 95% chance of falling in $\bar{x} \pm 1.96s/\sqrt{n}$ ”, although we are tempted to say this.

Thus, for a **frequentist**, “95%” is not really a coverage probability, but merely a tag associated with the interval to indicate how it would perform over the long haul.

A 99% frequentist interval would be wider, a 90% interval narrower, but conditional on the data, all would have coverage probability 0 or 1.

By contrast, Bayesian confidence intervals, known as **highest posterior density (HPD)** intervals (also **credible intervals**) do not require this awkward interpretation. In the Bayesian framework, we can make statements concerning the probability of θ falling in an interval, by specifying a prior distribution for θ . Thus, additional structure (i.e. a prior for θ) leads to an ease in interpretation.

2) Bayesian inference obeys the likelihood principle

The likelihood principle: If two distinct sampling plans (designs) yield proportional likelihood functions for θ , then inference about θ should be identical from these two designs. Frequentist inference does not obey the likelihood principle, in general.

Example 1.7

Suppose in 12 independent tosses of a coin, 9 heads and 3 tails are observed. I wish to test the null hypothesis $H_0 : \theta = 1/2$ vs. $H_a : \theta > 1/2$, where θ is the true probability of heads.

Consider the following 2 choices for the likelihood function:

- a) Binomial: Suppose $n = 12$ is **fixed** beforehand and let $x =$ number of heads in 12 tosses. Then $x \sim \text{Binomial}(12, \theta)$, and the likelihood function is given by

$$\begin{aligned} L_1(\theta) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \\ &= \binom{12}{9} \theta^9 (1 - \theta)^3 . \end{aligned}$$

- b) Negative Binomial: Suppose the experiment involved flipping the coin until the third tail appeared. Here, x = the number of heads required to complete the experiment, and so $x \sim \text{neg-binomial}(r = 3, \theta)$.

$$\begin{aligned} L_2(\theta) &= \binom{r+x-1}{x} \theta^x (1-\theta)^r \\ &= \binom{11}{9} \theta^9 (1-\theta)^3 . \end{aligned}$$

Note that $L_1(\theta) \propto L_2(\theta)$. From a Bayesian perspective, the posterior distribution of θ is the **same** under either design. That is

$$p(\theta | x) = \frac{L_1(\theta) \pi(\theta)}{\int L_1(\theta) \pi(\theta) d\theta} \equiv \frac{L_2(\theta) \pi(\theta)}{\int L_2(\theta) \pi(\theta) d\theta} .$$

However, under the frequentist paradigm, inferences about θ are quite different under each design. The rejection region is “reject H_0 if $X \geq c$ ”. Based on the binomial likelihood, the p-value is

$$\begin{aligned} p_1 &= P(x \geq 9 \mid \theta = 1/2) \\ &= \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = .075 , \end{aligned}$$

while for the negative binomial likelihood, the p-value is

$$\begin{aligned} p_2 &= P(x \geq 9 \mid \theta = 1/2) \\ &= \sum_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = .0325 . \end{aligned}$$

The two designs lead to different decisions. We would reject H_0 under the negative binomial design, but we would not reject H_0 under binomial sampling.

3) Bayesian inference does not lead to absurd results

Absurd results can be obtained when doing UMVUE estimation.

Suppose $x \sim \text{Poisson}(\lambda)$, and we want to estimate

$$\theta = \exp\{-2\lambda\}, \quad 0 < \theta < 1 .$$

It can be shown that the UMVUE of θ is $(-1)^x$. Thus, if x is even, the UMVUE of the θ is 1 and if x is odd, the UMVUE of θ is -1!

Such degenerate situations do not occur when inference is based on $p(\theta | x)$.

4) Bayes theorem is a formula for learning

Suppose you conduct an experiment and collect observations x_1, \dots, x_n .

Then

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

where $x = (x_1, \dots, x_n)$.

Suppose you collect an additional observation x_{n+1} in a new study.

Then

$$p(\theta | x, x_{n+1}) = \frac{p(x_{n+1} | \theta) p(\theta | x)}{\int p(x_{n+1} | \theta) p(\theta | x) d\theta} .$$

So your prior in the current study is the posterior from the previous study. Thus Bayes theorem updates information in a natural way.

5) Bayesian inference does not require large sample theory

With modern advances in computing, “exact” calculations can be carried out using Markov chain Monte Carlo (MCMC) methods. Bayes methods do not require asymptotics for valid inference. Thus small sample Bayesian inference proceeds in the same way as if one had a large sample.

6) Bayesian inference often has frequentist inference as a special case

Often one can obtain frequentist answers by choosing a uniform prior for the parameters, i.e., $\pi(\theta) \propto 1$, so that

$$p(\theta | x) \propto L(\theta) .$$

In such cases, frequentist answers can be obtained from such a posterior distribution.

Prior Distributions

There are various types of prior distributions that we need to discuss.

Noninformative Priors

Roughly speaking, a prior distribution is noninformative if the prior is “flat” relative to the likelihood function.

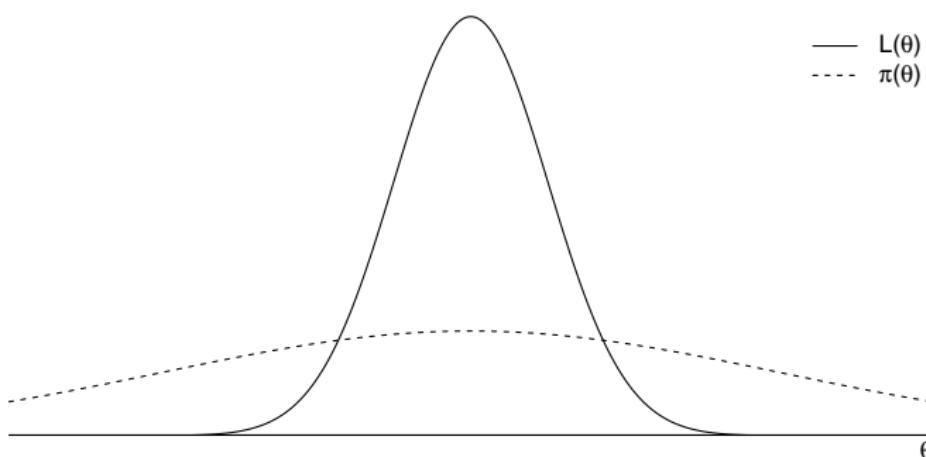


Figure 1.8:

Thus a prior $\pi(\theta)$ is noninformative if it has minimal impact on the posterior distribution of θ . There is a much more formal development of noninformative priors given in Box and Tiao (1973), Kass (1989, *Statistical Science*). Other names for noninformative prior are **reference prior**, **vague prior**, or **flat prior**.

In general a noninformative prior is one which is dominated by the likelihood, that is, it is a prior which does not change very much over the region in which the likelihood is appreciable and does not assume large values outside that range (see Figure 1.8). A prior which has these properties is said to be a **locally uniform** prior. Box and Tiao (1973) give a detailed development of local uniformity.

Examples of noninformative priors

1. If $0 \leq \theta \leq 1$, then $\theta \sim U(0, 1)$ is a noninformative prior for θ such that $\pi(\theta) = 1, 0 \leq \theta \leq 1$.

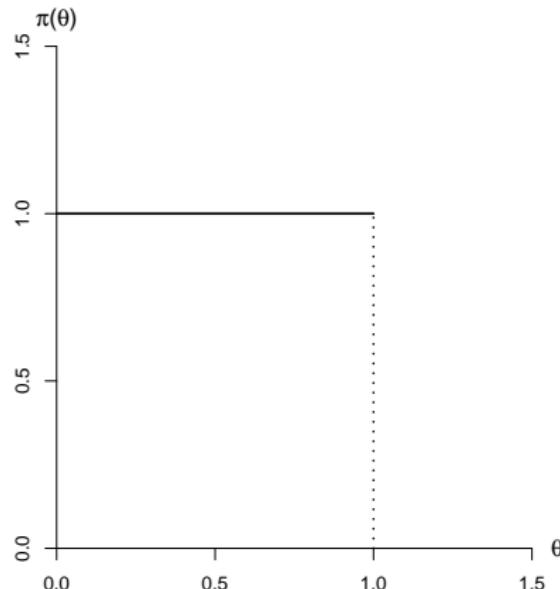


Figure 1.9:

2. If $-\infty < \theta < \infty$, then if $\theta \sim N(\mu_0, \sigma_0^2)$ and $\sigma_0^2 \rightarrow \infty$, then we get a noninformative prior. That is, we can pick σ_0^2 large enough so that we can obtain a noninformative prior.

Improper Priors

A prior $\pi(\theta)$ is said to be improper if

$$\int_{\Theta} \pi(\theta) d\theta = \infty .$$

Thus a prior is improper if its normalizing constant is equal to ∞ . Improper priors are often used in Bayesian inference since they usually yield noninformative priors.

Example 1.8

$-\infty < \theta < \infty$, $\pi(\theta) \propto 1$. That is θ has a uniform prior distribution on the real line. Clearly,

$$\int_{-\infty}^{\infty} \pi(\theta) d(\theta) = \int_{-\infty}^{\infty} d\theta = \infty .$$

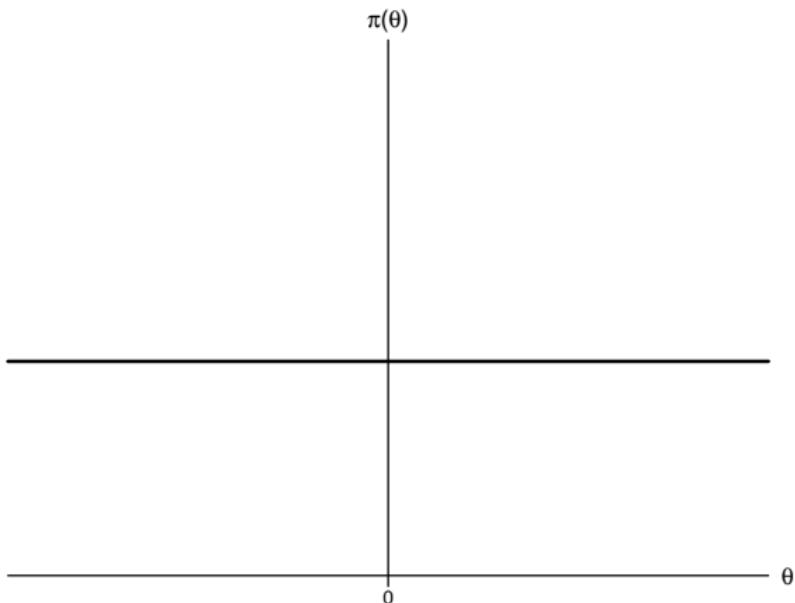


Figure 1.10:

Remark 1.6

1. An improper prior may result in an improper posterior. One **cannot** make inference with improper posterior distributions.
2. An improper prior may still lead to a **proper** posterior distribution.

Example 1.9

Suppose given θ , x_1, \dots, x_n are i.i.d. $N(\theta, 1)$. Here, $\Theta = \{\theta : -\infty < \theta < \infty\}$. Suppose $\pi(\theta) \propto 1$. Then the posterior distribution of θ is given by

$$\begin{aligned}
 p(\theta | x) &\propto p(x | \theta) \pi(\theta) \\
 &= \exp \left\{ -\frac{1}{2} \sum (x_i - \theta)^2 \right\} \times 1 \\
 &= \exp \left\{ -\frac{1}{2} \sum (x_i - \theta)^2 \right\} \\
 &\propto \exp \left\{ -\frac{1}{2} [n\theta^2 - 2\theta \sum x_i] \right\} \\
 &\propto \exp \left\{ -\frac{n}{2} [\theta - \bar{x}]^2 \right\}, \text{ where } \bar{x} = \frac{1}{n} \sum x_i .
 \end{aligned}$$

Thus $\theta | x \sim N(\bar{x}, \frac{1}{n})$.

Thus, the uniform improper prior on θ still leads to a proper posterior, i.e., a normal distribution with mean \bar{x} and variance $\frac{1}{n}$.

Example 1.10

Given θ , suppose x_1, \dots, x_n are i.i.d. $\text{Poisson}(\theta)$.

Suppose $\pi(\theta) \propto \theta^{-\frac{1}{2}}$. Here $\Theta = \{\theta : 0 < \theta < \infty\}$, and so

$$\int_0^\infty \pi(\theta) d\theta = \int_0^\infty \theta^{-\frac{1}{2}} d\theta = \infty.$$

Thus $\pi(\theta)$ is improper for θ .

$$\begin{aligned} p(\theta \mid x) &\propto (\theta^{\sum x_i} \exp\{-n\theta\}) \left(\theta^{-\frac{1}{2}}\right) \\ &= \theta^{\sum x_i - \frac{1}{2}} \exp\{-n\theta\} \\ &= \theta^{\sum x_i + \frac{1}{2} - 1} \exp\{-n\theta\} \end{aligned}$$

Thus $\theta \mid x \sim \text{gamma}\left(\frac{1}{2} + \sum x_i, n\right)$.

The posterior of θ is proper and is a gamma distribution.

Informative Priors

An informative prior is a prior **not dominated** by the likelihood, and has an impact on the posterior distribution.

If a prior $\pi(\theta)$ dominates the likelihood, it is clearly an informative prior. Informative priors must be specified with care in actual practice. They are reasonable priors to use if one has real prior information from a previous similar study, for example.

Example 1.11

Given θ , suppose x_1, \dots, x_{10} are i.i.d. $N(\theta, 10)$. Suppose $\theta \sim N(0, 1)$. Then this represents an informative prior for θ .

$$\begin{aligned} L(\theta) &= \exp \left\{ -\frac{n}{2(10)} (\theta - \bar{x})^2 \right\} \\ &= \exp \left\{ -\frac{10}{2(10)} (\theta - \bar{x})^2 \right\} \\ &= \exp \left\{ -\frac{1}{2} (\theta - \bar{x})^2 \right\}. \end{aligned}$$

$$\pi(\theta) \propto \exp \left\{ -\frac{1}{2} \theta^2 \right\}.$$

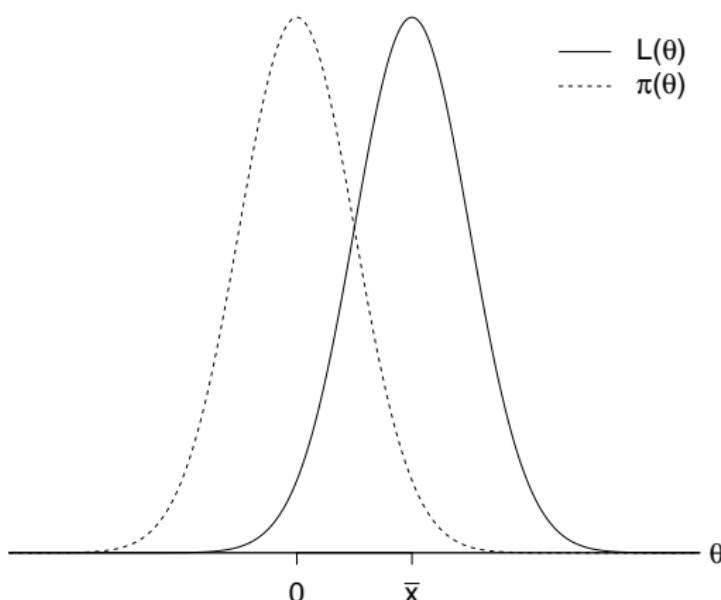


Figure 1.11: Informative priors

Example 1.12

Given θ , suppose x_1, \dots, x_{10} are i.i.d. $\text{Binomial}(1, \theta)$. Suppose that $\sum x_i = 5$, so that

$$\begin{aligned} p(x \mid \theta) &= \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \\ &= \theta^5 (1 - \theta)^5 \\ &= \theta^{6-1} (1 - \theta)^{6-1}. \end{aligned}$$

Now if $\theta \sim \text{beta}(5, 5)$, then this would be an informative prior for θ . In this case, $\pi(\theta) \propto \theta^{5-1} (1 - \theta)^{5-1}$.

Conjugate Priors

A prior is said to be a conjugate prior for a family of distributions if the prior and posterior distribution are of the same family.

Example 1.13

From earlier calculations, we showed that if x_1, \dots, x_n i.i.d. $\text{Binomial}(1, \theta)$, and $\theta \sim \text{beta}(\alpha, \lambda)$, then $\theta|x \sim \text{beta}(\alpha + \sum x_i, n - \sum x_i + \lambda)$. Thus, the beta prior is a conjugate prior for the binomial family.

Example 1.14

If x_1, \dots, x_n are i.i.d. $N(\mu, \sigma^2)$, σ^2 known, then the normal prior for μ is a conjugate prior.

<u>Family</u>	<u>Conjugate Prior</u>
Binomial(N, θ)	$\theta \sim \text{beta}(\alpha, \lambda)$
Poisson(θ)	$\theta \sim \text{gamma}(\delta_0, \gamma_0)$
$N(\mu, \sigma^2)$, σ^2 known	$\mu \sim N(\mu_0, \sigma_0^2)$
$N(\mu, \sigma^2)$, μ known, $\tau = 1/\sigma^2$	$\tau \sim \text{gamma}(\delta_0, \gamma_0)$
gamma(α, λ), α known	$\lambda \sim \text{gamma}(\delta_0, \gamma_0)$
Beta(α, λ), $\lambda = 1$	$\alpha \sim \text{gamma}(\delta_0, \gamma_0)$

Jeffreys Prior

Jeffreys prior is a prior that satisfies the local uniformity property for noninformative priors. It is a prior that is based on the Fisher information matrix.

Definition 1.2

Let $p(x | \theta)$ denote the density of x given θ . The Fisher information is

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

If θ is a $p \times 1$ vector, then

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

and thus $I(\theta)$ is a $p \times p$ matrix.

Definition 1.3

Jeffreys prior is defined as

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}},$$

where $|.|$ denotes determinant.

Theorem 1.3

Jeffreys prior is locally uniform and hence noninformative.

This property of Jeffreys prior is quite useful since it gives us an automated scheme for finding a noninformative prior for any parametric model $p(x | \theta)$.

Remark 1.7

Jeffreys prior is improper for many models. It may be proper, however, for certain models.

Example 1.15

Suppose x_1, \dots, x_n are i.i.d. Binomial($1, \theta$). Compute Jeffreys prior for θ .

The density based on one observation is

$$p(x | \theta) = \theta^x (1 - \theta)^{1-x} .$$

$$\log p(x | \theta) = x \log \theta + (1 - x) \log(1 - \theta) .$$

$$\frac{\partial}{\partial \theta} \log p(x | \theta) = \frac{x}{\theta} - \frac{(1 - x)}{1 - \theta} .$$

$$\frac{\partial^2}{\partial \theta^2} \log p(x | \theta) = -\frac{x}{\theta^2} - \frac{(1 - x)}{(1 - \theta)^2} .$$

$$\begin{aligned}
 I(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2} \log p(x \mid \theta)\right] \\
 &= \frac{E(x)}{\theta^2} + \frac{1-E(x)}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1-\theta}{(1-\theta)^2} \\
 &= \frac{1}{\theta} + \frac{1}{1-\theta} \\
 &= \frac{1}{\theta(1-\theta)} .
 \end{aligned}$$

Thus, Jeffreys prior is

$$\begin{aligned}
 \pi(\theta) &\propto I(\theta)^{\frac{1}{2}} \\
 &= \left(\frac{1}{\theta(1-\theta)}\right)^{\frac{1}{2}} \\
 &= \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} .
 \end{aligned}$$

Thus $\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} = \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1}$. Thus $\theta \sim \text{beta}(\frac{1}{2}, \frac{1}{2})$.

Therefore Jeffreys prior is **proper** for the binomial model.

Example 1.16

Suppose x_1, \dots, x_n are i.i.d. $\text{Poisson}(\theta)$. Find Jeffreys prior for θ .
Based on a single observation,

$$p(x | \theta) = \frac{\theta^x}{x!} \exp\{-\theta\}$$

$$\log p(x | \theta) = x \log \theta - \theta - \log(x!) .$$

$$\frac{\partial}{\partial \theta} \log p(x | \theta) = \frac{x}{\theta} - 1 .$$

$$\frac{\partial^2}{\partial \theta^2} \log p(x | \theta) = -\frac{x}{\theta^2} .$$

$$\begin{aligned} I(\theta) &= \frac{E(x)}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta} . \\ \pi(\theta) &\propto I(\theta)^{\frac{1}{2}} = \theta^{-\frac{1}{2}} . \end{aligned}$$

Thus Jeffreys prior is $\pi(\theta) \propto \theta^{-\frac{1}{2}}$, which is **improper**,
since $\int_0^\infty \pi(\theta) d\theta = \int_0^\infty \theta^{-\frac{1}{2}} d\theta = \infty$.

Example 1.17

Suppose x_1, \dots, x_n are i.i.d. $N(\mu, \sigma^2)$, where (μ, σ^2) are both unknown. Compute Jeffreys prior for (μ, σ) .

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

$$\log p(x | \mu, \sigma) = -\frac{1}{2} \log(2\pi) - \log \sigma - \frac{1}{2\sigma^2} (x - \mu)^2 .$$

$$\frac{\partial \log p(x | \mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} (x - \mu) .$$

$$\frac{\partial^2 \log p(x | \mu, \sigma)}{\partial \mu^2} = -\frac{1}{\sigma^2} .$$

$$\frac{\partial \log p(x | \mu, \sigma)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3} (x - \mu)^2 .$$

$$\frac{\partial^2 \log p(x | \mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3}{\sigma^4} (x - \mu)^2 .$$

$$\frac{\partial^2 \log p(x | \mu, \sigma^2)}{\partial \mu \partial \sigma} = -\frac{2}{\sigma^3} (x - \mu) .$$

Taking expectations, we get

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

$$\begin{aligned} \pi(\mu, \sigma) &\propto |I(\mu, \sigma)|^{\frac{1}{2}} \\ &= \left(\frac{1}{\sigma^2} \times \frac{2}{\sigma^2} \right)^{\frac{1}{2}} \\ &\propto \sigma^{-2}. \end{aligned}$$

Thus Jeffreys joint prior for (μ, σ) is

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

which is an improper prior.

An appealing property of Jeffreys prior is that it is **invariant** with respect to one-to-one transformations. This arises from the relation

$$I(\theta) = I(\psi(\theta)) \left(\frac{d\psi(\theta)}{d\theta} \right)^2$$

where $\psi(\theta)$ is a one-to-one function of θ . Thus

$$(I(\theta))^{1/2} = (I(\psi(\theta)))^{1/2} \left| \frac{d\psi(\theta)}{d\theta} \right| .$$

Note that $\left| \frac{d\psi(\theta)}{d\theta} \right|$ is the absolute value of the Jacobian of the transformation from θ to $\psi(\theta)$. Thus

$$(I(\theta))^{1/2} d\theta = (I(\psi))^{1/2} d\psi$$

Thus Jeffreys prior preserves scale in reparameterizations.

Example 1.18

Suppose $x \sim N(\mu, 1)$. Jeffreys prior for μ is $\pi(\mu) \propto 1$. Let $\psi(\mu) = \exp\{\mu\}$. This is a one-to-one transformation in μ . The corresponding Jeffreys prior for $\psi(\mu)$ is

$$\begin{aligned}(I(\psi(\mu)))^{1/2} &= (I(\mu))^{1/2} \left| \frac{d\psi(\mu)}{d\mu} \right|^{-1} \\ &= 1 \times \exp\{-\mu\} \\ &= \exp\{-\mu\}.\end{aligned}$$

Thus Jeffreys prior for $\gamma \equiv \psi(\mu) = \exp\{\mu\}$ is

$$\pi(\gamma) \propto \gamma^{-1}, \quad 0 < \gamma < \infty.$$

Remark 1.8

The invariance property of Jeffreys prior means that if we have a Jeffreys prior on θ (and hence locally uniform), and $\psi(\theta)$ is a one-to-one function of θ , then the Jeffreys prior for $\psi(\theta)$ is also a locally uniform prior for $\psi(\theta)$.

The multivariate version of this invariance principle is given by

$$|I(\theta)|^{\frac{1}{2}} = |I(\psi(\theta))|^{\frac{1}{2}} \left| \frac{\partial \psi(\theta)}{\partial \theta} \right|,$$

where

$$\frac{\partial \psi(\theta)}{\partial \theta} = \frac{\partial \psi_i(\theta)}{\partial \theta_j}, \quad i, j = l, \dots, p,$$

$\theta = (\theta_1, \dots, \theta_p)$, and $\psi(\theta) = (\psi_1(\theta), \dots, \psi_p(\theta))$.

Remark 1.9

If θ is vector valued, i.e., $\theta = (\theta_1, \dots, \theta_p)$, then Jeffreys prior for θ must be used with caution. Jeffreys prior in multiparameter settings sometimes gives strange results. As a result of this, modifications of Jeffreys prior have been proposed for multiparameter problems. (See Box and Tiao, 1973).

Predictive Distributions

Suppose an experiment is conducted in which $x = (x_1, \dots, x_n)$ is observed. The model is given by $p(x|\theta)$, with prior $\pi(\theta)$, and posterior

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta}.$$

Suppose we wish to predict a future observation z from this population. We wish to construct the **Bayesian predictive distribution** of z . This distribution is defined as $p(z|x)$.

We have

$$\begin{aligned}
 p(z|x) &= \frac{p(z,x)}{p(x)} = \frac{\int_{\Theta} p(z,x,\theta) d\theta}{\int_{\Theta} p(x,\theta) d\theta} \\
 &= \frac{\int_{\Theta} p(z,x|\theta)\pi(\theta) d\theta}{\int_{\Theta} p(x|\theta)\pi(\theta) d\theta} \\
 &= \frac{\int_{\Theta} p(z|\theta)p(x|\theta)\pi(\theta) d\theta}{\int_{\Theta} p(x|\theta)\pi(\theta) d\theta} \\
 &= \int_{\Theta} p(z|\theta) \left\{ \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta) d\theta} \right\} d\theta \\
 &= \int_{\Theta} p(z|\theta)p(\theta|x) d\theta .
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(z | x) &= \int_{\Theta} p(z | \theta) p(\theta | x) d\theta \\
 &= E_{\theta|x} [p(z | \theta)] .
 \end{aligned}$$

Here, $p(z | \theta)$ is the likelihood function of θ evaluated at z .

Example 1.19

Suppose x_1, \dots, x_n are i.i.d. $\text{Binomial}(1, \theta)$, and suppose $\theta \sim \text{beta}(\alpha, \lambda)$. Let us find the predictive distribution of a future observation z . We have

$$p(z | x) = \int_{\Theta} p(z | \theta) p(\theta | x) d\theta .$$

Now

$$p(z | \theta) = \theta^z (1 - \theta)^{1-z}, \quad z = 0, 1 ,$$

and

$$\begin{aligned} p(\theta | x) &\propto (\theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}) (\theta^{\alpha-1} (1 - \theta)^{\lambda-1}) \\ &= \theta^{\sum x_i + \alpha - 1} (1 - \theta)^{n - \sum x_i + \lambda - 1} . \end{aligned}$$

Thus,

$$\theta | x \sim \text{beta}(\sum x_i + \alpha, n - \sum x_i + \lambda) .$$

Now,

$$\begin{aligned}
 p(z|x) &= \int_0^1 \frac{\Gamma(n + \alpha + \lambda)}{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \lambda)} \theta^{z + \sum x_i + \alpha - 1} (1 - \theta)^{n+1 - \sum x_i + \lambda - z - 1} d\theta \\
 &= \left[\frac{\Gamma(n + \alpha + \lambda)}{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \lambda)} \right] \left[\frac{\Gamma(z + \sum x_i + \alpha)\Gamma(1 - z + n - \sum x_i + \lambda)}{\Gamma(n + \alpha + \lambda + 1)} \right].
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(z = 0 \mid x) &= \frac{\Gamma(n + \alpha + \lambda)\Gamma(n + 1 - \sum x_i + \lambda)}{\Gamma(n - \sum x_i + \lambda)\Gamma(n + \alpha + \lambda + 1)} \\
 &= \frac{n - \sum x_i + \lambda}{n + \alpha + \lambda},
 \end{aligned}$$

and

$$p(z = 1 \mid x) = \frac{\Sigma x_i + \alpha}{n + \alpha + \lambda} .$$

Note that $p(z = 1 \mid x) = E(\theta \mid x)$.

Example 1.20

Suppose x_1, \dots, x_n are i.i.d. $N(\theta, 1)$. Let z be a future observation. Assume $\pi(\theta) \propto 1$. Compute $p(z \mid x)$.

From an earlier computation, we have $\theta \mid x \sim N(\bar{x}, \frac{1}{n})$, where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Now,

$$p(z \mid \theta) = (2\pi)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(z - \theta)^2 \right\} .$$

Thus

$$\begin{aligned}
 p(z \mid x) &= \int_{-\infty}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(z-\theta)^2\right\} (2\pi)^{-\frac{1}{2}} \left(\frac{1}{n}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n}{2}(\theta - \bar{x})^2\right\} d\theta \\
 &\propto \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2} [\theta^2(1+n) - 2\theta(n\bar{x}+z)]\right\} \exp\left\{-\frac{z^2}{2}\right\} d\theta \\
 &\propto \exp\left\{-\frac{z^2}{2}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{(n+1)}{2} \left[\theta^2 - \frac{2\theta(n\bar{x}+z)}{n+1}\right]\right\} d\theta \\
 &\propto \exp\left\{\frac{n+1}{2} \left(\frac{n\bar{x}+z}{n+1}\right)^2\right\} \exp\left\{-\frac{z^2}{2}\right\} \\
 &\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{(n+1)}{2} \left[\theta - \frac{(n\bar{x}+z)}{n+1}\right]^2\right\} d\theta \\
 &\propto \exp\left\{-\frac{z^2}{2}\right\} \exp\left\{\frac{(n\bar{x}+z)^2}{2(n+1)}\right\} \\
 &\propto \exp\left\{-\frac{1}{2} \left[z^2 \left(\frac{n}{n+1}\right) - \frac{2n z\bar{x}}{n+1}\right]\right\} \\
 &\propto \exp\left\{-\frac{n}{2(n+1)} [z^2 - 2z\bar{x}]\right\} \\
 &\propto \exp\left\{-\frac{n}{2(n+1)} [z - \bar{x}]^2\right\}.
 \end{aligned}$$

Thus

$$\begin{aligned} z \mid x &\sim N\left(\bar{x}, \frac{n+1}{n}\right) \\ &= N\left(\bar{x}, 1 + \frac{1}{n}\right). \end{aligned}$$

Example 1.21

Suppose x_1, \dots, x_n are i.i.d. Poisson(θ). Suppose $\theta \sim \text{gamma}(\alpha, \lambda)$. Find $p(z \mid x)$.

$$\begin{aligned} p(\theta \mid x) &\propto \theta^{\sum x_i} \exp\{-n\theta\} \theta^{\alpha-1} \exp\{-\lambda\theta\} \\ &= \theta^{\sum x_i + \alpha - 1} \exp\{-\theta(\lambda + n)\}. \end{aligned}$$

Thus $\theta \mid x \sim \text{gamma}(\sum x_i + \alpha, \lambda + n)$.

Now

$$p(z \mid \theta) = \frac{\theta^z \exp\{-\theta\}}{z!}, \quad z = 0, 1, 2, \dots$$

Thus

$$\begin{aligned}
 p(z | x) &= \int_0^\infty \frac{\theta^z \exp\{-\theta\}}{z!} \frac{(\lambda + n)^{\sum x_i + \alpha}}{\Gamma(\sum x_i + \alpha)} \theta^{\sum x_i + \alpha - 1} \exp\{-\theta(\lambda + n)\} d\theta \\
 &= \frac{(\lambda + n)^{\sum x_i + \alpha}}{z! \Gamma(\sum x_i + \alpha)} \int_0^\infty \theta^{z + \sum x_i + \alpha - 1} \exp\{-\theta(\lambda + n + 1)\} d\theta \\
 &= \frac{(\lambda + n)^{\sum x_i + \alpha}}{z! \Gamma(\sum x_i + \alpha)} \frac{\Gamma(z + \sum x_i + \alpha)}{(\lambda + n + 1)^{z + \sum x_i + \alpha}} \\
 &= \binom{z + \sum x_i + \alpha - 1}{z} \left(\frac{\lambda + n}{\lambda + n + 1}\right)^{\sum x_i + \alpha} \left(\frac{1}{\lambda + n + 1}\right)^z, \quad z = 0, 1, 2, \dots
 \end{aligned}$$

Thus

$$z|x \sim \text{neg-binomial}\left(\sum x_i + \alpha, \frac{\lambda + n}{\lambda + n + 1}\right).$$

We note that z can also be a $q \times 1$ **vector** of future observations. In this case, the predictive distribution of z is a multivariate distribution.

Example 1.22

Suppose x_1, \dots, x_n are i.i.d. Binomial($1, \theta$), and $\theta \sim \text{beta}(\alpha, \lambda)$. Suppose we wish to construct the predictive distribution for a future vector $z = (z_1, \dots, z_q)$.

$$\begin{aligned} p(z | x) &= \left[\frac{\Gamma(n + \alpha + \lambda)}{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \lambda)} \right] \\ &\quad \times \left[\frac{\Gamma(\sum_{i=1}^q z_i + \sum x_i + \alpha) \Gamma(q - \sum_{i=1}^q z_i + n - \sum x_i + \lambda)}{\Gamma(n + \alpha + \lambda + q)} \right] \end{aligned}$$

Example 1.23

A study reported on the long-term effects of exposure to low levels of lead in childhood. Researchers analyzed children's shed primary teeth for lead content. Of the children those teeth had a lead content of more than 22.22 parts per million (ppm) (which is rather high), 22 eventually graduated from high school and 7 did not.

Suppose your prior density for the proportion of all such children who will graduate from high school is $\text{beta}(1,1)$, and so your posterior density is $\text{beta}(23,8)$.

Based on this information, of 10 more children who are found to have a lead content of more than 22.22 ppm, what is your (predictive) probability that 9 or 10 of them will graduate from high school? [Hint: Find the possibility of 9 and also of 10 and add them together.]

Here $\Sigma x_i = 22$, $n = 29$, $n - \Sigma x_i = 29 - 22 = 7$

$\alpha = 1$, $\lambda = 1$, $q = 10$, $E(\theta | x) = \frac{22+1}{29+1+1} = \frac{23}{31}$,
and $z = (z_1, \dots, z_{10})$.

Now $P(\text{nine graduate}) = \binom{10}{9} P(z_1=1, \dots, z_9=1, z_{10}=0)$.

$$\begin{aligned}
 p(z_1 = 1, \dots, z_9 = 1, z_{10} = 0) &= \left\{ \frac{\Gamma(29 + 1 + 1)}{\Gamma(22 + 1)\Gamma(29 - 22 + 1)} \right\} \left\{ \frac{\Gamma(9 + 22 + 1)\Gamma(10 - 9 + 29 - 22 + 1)}{\Gamma(29 + 1 + 1 + 10)} \right\} \\
 &= \left\{ \frac{\Gamma(31)}{\Gamma(23)\Gamma(8)} \right\} \left\{ \frac{\Gamma(32)\Gamma(9)}{\Gamma(41)} \right\} \\
 &= \left\{ \frac{\Gamma(32)}{\Gamma(23)} \right\} \left\{ \frac{\Gamma(31)}{\Gamma(41)} \right\} \left\{ \frac{\Gamma(9)}{\Gamma(8)} \right\} \\
 &= \left\{ \frac{(31)(30)(29)(28)(27)(26)(25)(24)(23)\Gamma(23)}{\Gamma(23)} \right\} \\
 &\quad \times \left\{ \frac{\Gamma(31)}{(40)(39)(38)(37)(36)(35)(34)(33)(32)(31)\Gamma(31)} \right\} \left\{ \frac{8\Gamma(8)}{\Gamma(8)} \right\} \\
 &= \left(\frac{23}{31} \right) \left(\frac{24}{32} \right) \left(\frac{25}{33} \right) \cdots \left(\frac{31}{39} \right) \left(\frac{8}{40} \right) = .0190 .
 \end{aligned}$$

Thus, $P(\text{nine graduate}) = 10(.0190) = .190$.

Note that if x is a positive integer, $\Gamma(x) = (x - 1)!$.

$$\begin{aligned} P(\text{all ten graduate}) &= P(z_1 = 1, \dots, z_{10} = 1) \\ &= \left(\frac{23}{31}\right) \left(\frac{24}{32}\right) \left(\frac{25}{33}\right) \cdots \left(\frac{32}{40}\right) \\ &= .076 . \end{aligned}$$

Thus, the desired probability is $.190 + .076 = .266$.

Example 1.24

The most important prognostic factor in early breast cancer is the number of axillary lymph nodes that test positive for breast cancer, indicating the extent to which the cancer has spread. There is no standard number of lymph nodes to sample. Sometimes surgeons remove three and sometimes they remove 30.

The proportion of lymph nodes that are positive has approximately a beta(.1,5) density. We can calculate the marginal probability that the first lymph node sampled is positive:

$$\begin{aligned}
 P(x_1 = 1) &= \int_0^1 P(x_1 = 1 | \theta) \pi(\theta) d\theta \\
 &= \frac{1}{B(.1, 5)} \int_0^1 \theta^{.1-1} (1-\theta)^{5-1} d\theta \\
 &= E[\text{beta}(.1, 5)] \\
 &= \frac{.1}{.1 + 5} \\
 &\approx .02 .
 \end{aligned}$$

So the probability that the first lymph node sampled is positive is $.1/5.1$ or about 2%. But if the first one sampled tests positive, the updated density is beta(1.1,5) and so the conditional probability that the second is positive is $1.1/6.1$ or about 18%.

A woman with breast cancer had a mastectomy and, subsequently, the surgeon removed three lymph nodes. None tested positive. Another doctor questions the surgeon, suggesting that had more been removed, perhaps some would have been positive.

Should the surgeon have removed more? To address this, suppose the surgeon had removed an additional 20 lymph nodes (for a total of 23).

What is the probability that none of the 23 would have tested positive?
(Assume that the 23 are exchangeable initially.)

First, consider the hypothetical case if only one lymph node were to be removed and it tested positive. Let

$$\begin{aligned}\theta &= \text{proportion of positive lymph notes} \\ \theta &\sim \text{beta}(.1, 5) \\ x_1 &= 1.\end{aligned}$$

$$\begin{aligned}p(\theta|x_1) &\propto p(x_1|\theta)\pi(\theta) \\ &\propto \theta^{x_1}(1-\theta)^{1-x_1}\theta^{.1-1}(1-\theta)^{5-1} \\ &= \theta^{x_1+.1-1}(1-\theta)^{6-x_1-1} \\ &= \theta^{1.1-1}(1-\theta)^{5-1}.\end{aligned}$$

Thus, $\theta|x_1 \sim \text{beta}(1.1, 5)$.

$$\begin{aligned}p(z|x_1 = 1) &= E_{\theta|x} [f(z|\theta)] \\ &= E_{\theta|x} [\theta^z(1-\theta)^{1-z}] \\ p(z = 1|x_1 = 1) &= E_{\theta|x} [\theta] = \frac{1.1}{1.1 + 5} = .18.\end{aligned}$$

Actual experiment:

$$x_1 = 0, x_2 = 0, x_3 = 0, x = (x_1, x_2, x_3) = (0, 0, 0), z = (z_1, \dots, z_{20})$$

$$\begin{aligned}\pi(\theta) &\propto \theta^{.1-1} (1-\theta)^{5-1} \\ p(\theta|x) &\propto \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \theta^{.1-1} (1-\theta)^{5-1} \\ &= \theta^{\sum_{i=1}^n x_i + .1 - 1} (1-\theta)^{n - \sum_{i=1}^n x_i + 5 - 1} \\ &= \theta^{.1-1} (1-\theta)^{3-0+5-1} \\ &= \theta^{.1-1} (1-\theta)^{8-1} \\ \theta|x &\sim \text{beta}(.1, 8).\end{aligned}$$

Thus, the posterior density is $\text{beta}(.1, 5+3) = \text{beta}(.1, 8)$.

From page 109,

$$\begin{aligned}p(z_1 = 0, \dots, z_{20} = 0|x) &= \left\{ \frac{\Gamma(n + \alpha + \lambda)}{\Gamma(\sum_{i=1}^n x_i + \alpha) \Gamma(n - \sum_{i=1}^n x_i + \lambda)} \right\} \\ &\times \left\{ \frac{\Gamma(\sum_{i=1}^{20} z_i + \sum_{i=1}^n x_i + \alpha) \Gamma(20 - \sum_{i=1}^{20} z_i + n - \sum_{i=1}^n x_i + \lambda)}{\Gamma(n + \alpha + \lambda + 20)} \right\}\end{aligned}$$

$$n = 3, \alpha = .1, \lambda = 5, q = 20, \sum_{i=1}^n x_i = 0, \sum_{i=1}^q z_i = 0.$$

The predictive probability that the additional 20 lymph nodes would all be negative is:

$$\begin{aligned}
 p(z_1 = 0, \dots, z_{20} = 0 | x) &= \left\{ \frac{\Gamma(3 + .1 + 5)}{\Gamma(.1)\Gamma(3 + 5)} \right\} \left\{ \frac{\Gamma(.1)\Gamma(23 + 5)}{\Gamma(3 + .1 + 5 + 20)} \right\} \\
 &= \frac{\Gamma(8.1)}{\Gamma(.1)\Gamma(8)} \frac{\Gamma(.1)\Gamma(28)}{\Gamma(28.1)} \\
 &= \frac{\Gamma(8.1)}{\Gamma(28.1)} \frac{\Gamma(28)}{\Gamma(8)} .
 \end{aligned}$$

Note that

$$\begin{aligned}
 \Gamma(28.1) &= \Gamma(27.1 + 1) = 27.1 \Gamma(27.1) \\
 &= 27.1 \Gamma(26.1 + 1) \\
 &= (27.1)(26.1) \Gamma(26.1) .
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(z_1 = 0, \dots, z_{20} = 0 | x) &= \frac{\Gamma(8.1)}{\Gamma(28.1)} \frac{\Gamma(28)}{\Gamma(8)} \\
 &= \frac{\Gamma(8.1)}{(27.1)(26.1)(25.1) \cdots (8.1)\Gamma(8.1)} \left\{ \frac{27 \cdot 26 \cdots 8\Gamma(8)}{\Gamma(8)} \right\} \\
 &= \left(\frac{27}{27.1} \right) \left(\frac{26}{26.1} \right) \cdots \left(\frac{8}{8.1} \right) \\
 &= \left(\frac{8}{8.1} \right) \left(\frac{8+1}{8.1+1} \right) \left(\frac{8+2}{8.1+2} \right) \cdots \left(\frac{8+19}{8.1+19} \right) = .8786 .
 \end{aligned}$$

Standard distributions and their conjugate priors			
Notation and name	Random variable and domain	Probability density function	Parameter restrictions [Vague prior information]
1 Binomial			
Binomial(n, θ)	x $x = 0, 1, \dots, n$	$\binom{n}{x} \theta^x (1 - \theta)^{n-x}$	$0 \leq \theta \leq 1$
Binomial			
Beta(g, h)	θ $0 \leq \theta \leq 1$	$\frac{\theta^{g-1} (1-\theta)^{h-1}}{B(g,h)}$	$g > 0, h > 0$ $[g \rightarrow 0, h \rightarrow 0]$
Beta			
2 Poisson			
Poisson(θ)	x $x = 0, 1, 2, \dots$	$\frac{\theta^x \exp\{-\theta\}}{x!}$	$\theta > 0$
Poisson			
Gamma(g, h)	θ $\theta > 0$	$\frac{h^g \theta^{g-1} \exp\{-h\theta\}}{\Gamma(g)}$	$g > 0, h > 0$ $[g \rightarrow 0, h \rightarrow 0]$
Gamma			
3 Gamma			
Gamma(k, θ)	x $x > 0$	$\frac{\theta^k x^{k-1} \exp\{-\theta x\}}{\Gamma(k)}$	$\theta > 0, k > 0$
Gamma			
Gamma(g, h)	θ $\theta > 0$	$\frac{h^g \theta^{g-1} \exp\{-h\theta\}}{\Gamma(g)}$	$g > 0, h > 0$ $[g \rightarrow 0, h \rightarrow 0]$
Gamma			
Exponential(θ) = Gamma($1, \theta$)	x $x > 0$	$\theta \exp\{-\theta x\}$	$\theta > 0$
Exponential			

Standard distributions and their conjugate priors (continued)

Notation and name	Random variable and domain	Probability density function	Parameter restrictions [Vague prior information]
4 Multinomial			
Mu(n, θ)	$x = (x_1, \dots, x_d)$		
Multinomial	$x_i = 0, 1, \dots, n,$ $\sum x_i \leq n$	$\binom{n}{x} \theta_1^{x_1} \dots \theta_d^{x_d} (1 - \sum \theta_i)^{n - \sum x_i}$	$0 \leq \theta_i \leq 1$ $\sum \theta_i \leq 1$
Di(g, h)	$\theta = (\theta_1, \dots, \theta_d)$		
Dirichlet	$0 \leq \theta_i \leq 1,$ $\sum \theta_i \leq 1$	$\frac{\theta_1^{g_1-1} \dots \theta_d^{g_d-1} (1 - \sum \theta_i)^{h-1}}{D(g, h)}$	$g > 0, h > 0$ $[g \rightarrow 0, h \rightarrow 0]$
5 Normal			
$N(\mu, \sigma^2), \tau = \frac{1}{\sigma^2}$	x		
Normal	$x \in \mathbb{R}^1$	$\left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}\tau(x - \mu)^2\right\}$	$\mu \in \mathbb{R}^1, \tau > 0$
NoCh(b, c, g, h)	(μ, τ)	$p(\mu \tau) \text{ is } N(b, c\tau)$	
Normal chi-squared	$\mu \in \mathbb{R}^1, \tau > 0$	$p(\tau) \text{ is Ch}(g, h):$ $\frac{\left(\frac{1}{2}h\right)^{g/2} \tau^{(g/2)-1} \exp\left\{-\frac{1}{2}h\tau\right\}}{\Gamma\left(\frac{1}{2}g\right)}$	$b \in \mathbb{R}^1, c > 0$ $g > 0, h > 0$ $[c \rightarrow 0, g \rightarrow 0, h \rightarrow 0]$
6 Multinormal			
$N_d(\mu, \sigma^2), \tau = \frac{1}{\sigma^2}$	$x = (x_1, \dots, x_d)$		
Normal	$x \in \mathbb{R}^d$	$\frac{ \tau ^{1/2}}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}(x - \mu)' \tau (x - \mu)\right\}$	$\mu \in \mathbb{R}^d$ $\tau \in \delta^d$
NoWi _d (b, c, g, h)	(μ, τ)	$p(\mu \tau) \text{ is } N_d(b, c\tau)$	
Normal-Wishart	$\mu \in \mathbb{R}^d, \tau \in \delta^d$	$p(\tau) \text{ is Wi}_d(g, h):$ $\frac{ \frac{1}{2}h ^{g/2} \tau ^{(g-d-1)/2} \exp\left\{-\frac{1}{2}\text{tr}h\tau\right\}}{\Gamma_d\left(\frac{1}{2}g\right)}$	$b \in \mathbb{R}^d, c > 0$ $g > d - 1, h \in \delta^d$ $[c \rightarrow 0, g \rightarrow 0, h \rightarrow 0]$

Standard distributions and their conjugate priors (continued)				
Notation and name	Random variable and domain	Probability density function	Parameter restrictions [Vague prior information]	
7 Exponential (two-parameter) El(μ, τ) Exponential left-sided	x $x < \mu$	$\tau \exp\{-\tau(\mu - x)\}$	$\mu \in \mathbb{R}^1, \tau > 0$	
ErGamma(b, c, g, h) Exponential (right-sided)-gamma	(μ, τ) $\mu \in \mathbb{R}^1, \tau > 0$	$p(\mu \tau)$ is Er($b, c\tau$) $p(\tau)$ is Gamma(g, h)	$b \in \mathbb{R}^1, c > 0$ $g > 0, h > 0$ $[b \rightarrow -\infty, c \rightarrow 0,$ $g \rightarrow 0, h \rightarrow 0]$	
Er(μ, τ) Exponential right-sided	x $x \geq \mu$	$\tau \exp\{-\tau(x - \mu)\}$	$\mu \in \mathbb{R}^1, \tau > 0$	
ElGamma(b, c, g, h) Exponential (left-sided)-gamma	(μ, τ) $\mu \in \mathbb{R}^1, \tau > 0$	$p(\mu \tau)$ is El($b, c\tau$) $p(\tau)$ is Gamma(g, h)	$b \in \mathbb{R}^1, c > 0$ $g > 0, h > 0$ $[b \rightarrow \infty, c \rightarrow 0,$ $g \rightarrow 0, h \rightarrow 0]$	

Predictive density functions			
Notation and name	Random variable and domain	Probability density function	Parameter restrictions
1. BeBi(n, g, h) Beta-binomial	y $y = 0, 1, \dots, n$	$\binom{n}{y} \frac{B(g+y, h+n-y)}{B(g, h)}$	n positive integer, $g > 0, h > 0.$
2. NB(n, l) Negative-binomial	y $y = 0, 1, 2, \dots$	$\binom{y+n-1}{n-1} l^y (1-l)^n$	n positive integer, $0 \leq l \leq 1.$
3. InBe(k, g, h) Inverse-beta	y $y > 0$	$\frac{h^g y^{k-1}}{B(k, g)(h+y)^{g+k}}$	$k > 0, g > 0, h > 0.$
4. DiMu(k, g, h) Dirichlet-multinomial	$y = (y_1, \dots, y_d)$ $y_i = 0, 1, \dots, n$ $\sum y_i \leq n$	$\binom{n}{y} \frac{D(g+y, h+n-\sum_i^d y_i)}{D(g, h)}$	$g = (g_1, \dots, g_d),$ $g_i \geq 0,$ $h > 0.$
5. St(k, b, c) Student	m $m \in \mathbb{R}^1$	$\frac{\{1+(kc)^{-1}(m-b)^2\}^{-(k+1)/2}}{B(\frac{1}{2}, \frac{1}{2}k)(kc)^{1/2}}$	$b \in \mathbb{R}^1,$ $c > 0, k > 0.$
Si(k, g, h) Siegel	ν $\nu > 0$	$\frac{\nu^{(g/2)-1} (1+h^{-1}\nu)^{-(k+g)/2}}{B(\frac{1}{2}k, \frac{1}{2}g) h^{g/2}}$	$k > 0,$ $g > 0, h > 0.$
StSi($k; b, c; g, h$) Student-Siegel	(m, ν) $m \in \mathbb{R}^1, \nu > 0$	$\frac{\{1+(kc)^{-1}(m-b)^2 + h^{-1}\nu\}^{-(k+g+1)/2}}{D(\frac{1}{2}, \frac{1}{2}k, \frac{1}{2}g)(kc)^{1/2} h^{g/2} \nu^{1-(g/2)}}$	$b \in \mathbb{R}^1, c > 0,$ $k > 0, g > 0, h > 0.$

Predictive density functions (continued)			
Notation and name	Random variable and domain	Probability density function	Parameter restrictions
6. $\text{St}_d(k, b, c)$ Student	$m \in \mathbb{R}^d$	$\frac{\{1+(m-b)'(kc)^{-1}(m-b)\}^{-(k+1)/2}}{D\left(\frac{1}{2}\mathbf{1}, \frac{1}{2}(k-d+1)\right) kc ^{1/2}}$	$b \in \mathbb{R}^d, c \in \delta^d,$ $k > d - 1.$
$\text{Si}_d(k, g, h)$ Siegel	$\nu \in \delta^d$	$\frac{ \nu ^{(g-d-1)/2}}{B_g\left(\frac{1}{2}k, \frac{1}{2}g\right) h ^{g/2} I+h^{-1}\nu ^{(k+g)/2}}$	$k > d - 1,$ $g > d - 1,$ $h \in \delta^d.$
$\text{StSi}_d(k; b, c; g, h)$ Student-Siegel	$(m, \nu) \quad m \in \mathbb{R}^d, \nu \in \delta^d$	$\frac{\Gamma_d\left\{\frac{1}{2}(k+g+1)\right\} h ^{-g/2} \nu ^{(g-d-1)/2}}{f(k, b, c, g, h)}$	$k > d - 1, b \in \mathbb{R}^d,$ $c \in \delta^d, g > d - 1,$ $h \in \delta^d.$
where $f(k, b, c, g, h) = \Gamma_d\left(\frac{1}{2}k\right)\Gamma_d\left(\frac{1}{2}g\right)\pi^{d/2} kc ^{1/2} I + (kc)^{-1}(m-b)(m-b)' + h^{-1}\nu ^{(k+g+1)/2}$			

Construction of predictive distributions

$p(x \theta)$	$p(\theta)$
$p(y \theta)$	$p(\theta x)$
$p(y x)$	
<hr/>	
1 Binomial	
Binomial(n, θ)	Beta(g, h)
Binomial(N, θ)	Beta(G, H)
BetaBinomial(N, G, H)	where $G = g + x$, $H = h + n - x$.
<hr/>	
2 Poisson	
Poisson($k\theta$)	Gamma(g, h)
Poisson($K\theta$)	Gamma(G, H)
NB($G, \frac{K}{K+H}$)	where $G = g + x$, $H = h + k$.
<hr/>	
3 Gamma	
Gamma(k, θ)	Gamma(g, h)
Gamma(K, θ)	Gamma(G, H)
InverseBeta(K, G, H)	where $G = g + k$, $H = h + x$.
<hr/>	
4 Multinomial	
Mu(n, θ)	Dirichlet(g, h)
Mu(N, θ)	Dirichlet(G, H)
DiMu(N, G, H)	where $G = g + x$, $H = h + n - \mathbf{1}'x$.

Construction of predictive distributions (continued)

$p(x \theta)$	$p(\theta)$
$p(y \theta)$	$p(\theta x)$
$p(y x)$	

5 Normal

$$\begin{cases} N(\mu, k^{-1}\sigma^2), \tau = \frac{1}{\sigma^2} \\ \text{Ch}(\nu, \tau) \end{cases} \quad \text{NoCh}(b, c, g, h)$$

$$\begin{cases} N(\mu, K^{-1}\sigma^2), \tau = \frac{1}{\sigma^2} \\ \text{Ch}(\lambda, \tau) \end{cases} \quad \text{NoCh}(B, C, G, H) \quad \text{where } B = C^{-1}(cb + km), \\ C = c + k,$$

$$\begin{cases} \text{St} \left\{ G, B, \left(\frac{1}{K} + \frac{1}{C} \right) \frac{H}{G} \right\} \\ \text{Si}(G, \lambda, H) \\ \text{StSi} \left\{ G; B, \left(\frac{1}{K} + \frac{1}{C} \right) \frac{H}{G}; \lambda, H \right\} \end{cases} \quad G = g + \nu + \Delta(c), \\ H = h + \nu + \frac{ck}{c+k}(m - b)^2. \\ \Delta(c) = \begin{cases} 0 & (c = 0), \\ 1 & (c > 0). \end{cases}$$

6 Multinormal

$$\begin{cases} N_d(\mu, k^{-1}\sigma^2), \tau = \frac{1}{\sigma^2} \\ \text{Wi}_d(\nu, \tau) \end{cases} \quad \text{NoWi}_d(b, c, g, h)$$

$$\begin{cases} N_d(\mu, K^{-1}\sigma^2), \tau = \frac{1}{\sigma^2} \\ \text{Wi}_d(\lambda, \tau) \end{cases} \quad \text{NoWi}_d(B, C, G, H) \quad \text{where } B = C^{-1}(cb + km), \\ C = c + k, \\ G = g + \nu + \Delta(c), \\ H = h + \nu + \frac{kc}{c+k}(m - b)(m - b)'$$

Construction of predictive distributions (continued)

$p(x \theta)$	$p(\theta)$
$p(y \theta)$	$p(\theta x)$
$p(y x)$	
7 Exponential (two-parameter)	
$\begin{cases} \text{Er}(\mu, k\tau) \\ \text{Gamma}(\nu, \tau) \end{cases}$	$\text{ElGa}(b, c, g, h)$
$\begin{cases} \text{Er}(\mu, K\tau) \\ \text{Gamma}(\lambda, \tau) \end{cases}$	$\text{ElGa}(B, C, G, H)$ where $B = \min(m, b)$, $C = c + k$, $G = g + \nu + \Delta(c)$, $H = h + \nu + \omega_m(b, c, k)(m - b)$. $\omega_m(b, c, k) = \begin{cases} -c & (m < b) \\ k & (m \geq b) \end{cases}$.
$\begin{cases} \frac{CK}{B(1,G)(C+K)H} \left\{ 1 + H^{-1}\omega_M(B, C, K)(M - B) \right\}^{-(G+1)} \\ \text{InBe}(\lambda, G, H) \\ \frac{CKV^{\lambda-1}}{D(1,G,\lambda)(C+K)H^{\lambda+1}} \left\{ 1 + H^{-1}\omega_M(B, C, K)(M - B) + H^{-1}V \right\}^{-(G+\lambda+1)} \end{cases}$	

Bayesian Hypothesis Testing

The quantity used to test hypotheses in the Bayesian framework is called the **Bayes factor**. We introduced the Bayes factor earlier in the example involving paternity suits.

Suppose we wish to test the hypothesis of H_0 versus H_1 . As Bayesians, we specify prior probabilities for H_0 and H_1 . We denote these $p(H_0)$ and $p(H_1)$. Let D denote the data collected in the experiment, and let $p(H_0 | D)$ and $p(H_1 | D)$ denote the posterior probabilities of H_0 and H_1 , respectively.

Then the Bayes factor in **favor** of H_0 is defined as posterior to prior odds of H_0 divided by the posterior to prior odds to H_1 .

That is,

$$\begin{aligned}
 BF &= \frac{p(H_0 | D)/p(H_0)}{p(H_1 | D)/p(H_1)} \\
 &= \frac{\left(\frac{p(H_0|D)}{p(H_1|D)}\right)}{\left(\frac{p(H_0)}{p(H_1)}\right)}. \tag{1.3}
 \end{aligned}$$

Now by Bayes theorem, we know that

$$p(H_0 | D) = \frac{p(D | H_0) p(H_0)}{p(D | H_0)p(H_0) + p(D | H_0^c)p(H_0^c)}.$$

A similar formula holds for $p(H_1 | D)$. Substituting these into (1.3), we get the Bayes factor in favor of H_0 is

$$BF = \frac{p(D | H_0)}{p(D | H_1)}. \tag{1.4}$$

The Bayes factor has a very nice interpretation. Large values of BF in (1.3) favor H_0 . The following table was devised by Jeffreys to classify the strength of evidence in favor of H_0 :

$1 \leq BF \leq 3$	Weak
$3 < BF \leq 12$	Positive
$12 < BF \leq 150$	Strong
$BF > 150$	Decisive

More formally, suppose our parameter space is Θ , and we wish to test

$$\begin{aligned} H_0 : \theta &\in \Theta_{H_0} \\ H_1 : \theta &\in \Theta_{H_1}, \end{aligned}$$

where

$$\Theta = \Theta_{H_0} \cup \Theta_{H_1} \quad \text{and} \quad \Theta_{H_0} \cap \Theta_{H_1} = \emptyset.$$

Then, the Bayes factor in favor of H_0 is defined as

$$BF = \frac{p(D | H_0)}{p(D | H_1)} = \frac{\int_{\Theta_{H_0}} p(D | \theta, H_0) \pi(\theta | H_0) d\theta}{\int_{\Theta_{H_1}} p(D | \theta, H_1) \pi(\theta | H_1) d\theta} ,$$

where Θ_{H_0} is the parameter space under H_0 , and Θ_{H_1} is the parameter space under H_1 .

One immediate feature of the Bayes factor is that $p(D | H_0)$ and $p(D | H_1)$ are obtained by integrating over the parameter space, **NOT** maximizing over it.

The Bayes factor is the Bayesian analog of the likelihood ratio test for frequentists. It has several advantages over the likelihood ratio test.

1. We integrate over the parameter space instead of maximize over it.
2. The BF does **NOT** require nested models.
3. The BF has a nice interpretation of posterior to prior odds ratio.
4. The BF is a monotonic function of the posterior probability when $P(H_0) = P(H_1) = .5$.
5. The BF reduces to the likelihood ratio test in the case of a simple vs. simple hypothesis test.

For

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1 ,$$

$$BF = \frac{p(x | \theta = \theta_0)}{p(x | \theta = \theta_1)} .$$

For general hypotheses, the BF expressed in terms of densities is given by

$$BF = \frac{\int_{\Theta_{H_0}} p(x | \theta, H_0) \pi(\theta | H_0) d\theta}{\int_{\Theta_{H_1}} p(x | \theta, H_1) \pi(\theta | H_1) d\theta} .$$

Remark 1.10

1. BF is defined **only** when **proper** prior distributions are defined for θ . The BF is **NOT** defined for improper prior distributions since $p(x)$ is always improper when the prior distribution for θ is improper.
2. The BF may be sensitive to the choice of prior distribution and/or the choice of prior hyperparameters. For example, if $\theta \sim N(\mu_0, \sigma_0^2)$, the BF may become sensitive as $\sigma_0^2 \rightarrow \infty$.

Example 1.25: Simple vs. Simple

Suppose x_1, \dots, x_n are i.i.d. $N(\theta, 1)$ and we wish to test

$$\begin{aligned} H_0 : \quad & \theta = 0 \\ H_1 : \quad & \theta = 1 . \end{aligned}$$

Then

$$\begin{aligned} BF &= \frac{(2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum x_i^2\right\}}{(2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum(x_i - 1)^2\right\}} \\ &= \exp\left\{\frac{n}{2} - \sum x_i\right\} . \end{aligned}$$

Suppose $n = 10$ and $\sum_{i=1}^{10} x_i = 4.5$. Then

$$BF = \exp\left\{\frac{10}{2} - 4.5\right\} = \exp\{.5\} = 1.65 ,$$

which is weak evidence in favor of $H_0 : \theta = 0$.

Now if $\sum x_i = 1$, then

$$BF = \exp\{5 - 1\} = \exp\{4\} = 55 ,$$

which is strong evidence in favor of $H_0 : \theta = 0$.

Example 1.26: Simple vs. Composite

Suppose we wish to test

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0 .$$

$$BF = \frac{p(x | \theta = \theta_0)}{\int_{\Theta_{H_1}} p(x | \theta) \pi(\theta) d\theta} .$$

For the normal example, if we choose $\theta_0 = 0$, and $\pi(\theta) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - 1)^2\right\}$, then

$$\begin{aligned}
 BF &= \frac{(2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum x_i^2\right\}}{\int_{-\infty}^{\infty} (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum(x_i - \theta)^2\right\} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - 1)^2\right\} d\theta} \\
 &= \frac{(2\pi)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum x_i^2\right\}}{\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(\sum x_i^2 - 2\theta \sum x_i + n\theta^2)\right\} \exp\left\{-\frac{1}{2}(\theta^2 - 2\theta + 1)\right\} d\theta} \\
 &= \frac{(2\pi)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum x_i^2\right\}}{\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(1 + \sum x_i^2)\right\} \exp\left\{\frac{(1 + \sum x_i)^2}{2(n+1)}\right\} \exp\left\{-\frac{(n+1)}{2}(\theta - \frac{(1 + \sum x_i)}{n+1})^2\right\} d\theta} \\
 &= \frac{(2\pi)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum x_i^2\right\}}{(2\pi)^{\frac{1}{2}}(n+1)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(1 + \sum x_i^2)\right\} \exp\left\{\frac{(1 + \sum x_i)^2}{2(n+1)}\right\}} \\
 &= (n+1)^{\frac{1}{2}} \exp\left\{\frac{1}{2}\right\} \exp\left\{-\frac{(1 + \sum x_i)^2}{2(n+1)}\right\}.
 \end{aligned}$$

If $n = 10$ and $\sum x_i = 5$, we get $BF = 1.06$.

Posterior Model Probabilities

An alternative to the Bayes factor is the posterior model probability. Suppose we have k models denoted by m_1, \dots, m_k , and we want to select the “best” model. Our selection criterion is based on the model with the largest posterior probability. By Bayes theorem, we have

$$p(m_j | D) = \frac{p(D | m_j) p(m_j)}{\sum_{i=1}^k p(D | m_i) p(m_i)} .$$

The quantity $p(m_j)$ is the prior probability of model m_j , and $p(D|m_j)$ is the marginal distribution of the data given model m_j .

That is

$$p(D|m_j) = \int_{\Theta_{m_j}} p(x|\theta, m_j) \pi(\theta|m_j) d\theta ,$$

where Θ_{m_j} denotes the parameter space under model m_j . Note, that unlike the Bayes factor, $p(m_j|D)$ depends on the prior model probabilities $p(m_j)$.

The relationship between Bayes factors and posterior model probabilities is given by

$$p(m_j|D) = \frac{\alpha_j BF_{j1}}{\sum_{j=1}^k \alpha_j BF_{j1}} ,$$

where $\alpha_j = \frac{p(m_j)}{p(m_1)}$, and BF_{j1} is the Bayes factor in favor of m_j over m_1 , $j = 1, \dots, k$.

That is

$$BF_{j1} = \frac{\int_{\Theta_{m_j}} p(x|\theta, m_j) \pi(\theta|m_j) d\theta}{\int_{\Theta_{m_1}} p(x|\theta, m_1) \pi(\theta|m_1) d\theta}.$$

Note that if the prior model probabilities are all equal, i.e., $p(m_1) = p(m_2) = \dots = p(m_k) = \frac{1}{k}$, then the Bayes factor is a monotonic function of the posterior model probability, and thus the two methods are equivalent in this case. This is clear since in this case, $\alpha_j = 1$ and

$$p(m_j|D) = \frac{BF_{j1}}{\sum_{j=1}^k BF_{j1}}.$$

Let $\mathcal{M} = \{m_1, \dots, m_k\}$ denote the collection of models. \mathcal{M} is often referred to as the **model space**. In most model selection problems, such as variable subset selection, the model space is discrete and finite.

In model selection problems, in addition to specifying prior distributions for all of the parameters, we must also specify a discrete prior distribution for the model space \mathcal{M} . A uniform prior on \mathcal{M} corresponds to $p(m_j) = \frac{1}{k}$, $j = 1, \dots, k$.

A fully Bayesian approach to model selection is to compute posterior model probabilities for all possible models in the model space, and to select the model with the largest posterior probability.

There are also several **criterion based** approaches to model selection that have a Bayesian motivation. A common criterion for model selection is called the BIC criterion (Schwartz, 1978, *Annals of Statistics*). BIC has a Bayesian motivation, and is given by

$$\text{BIC}_j = -2 \log L(\hat{\theta}|m_j) + \log(n)p_j, \quad j = 1, \dots, k,$$

where p_j = dimension of θ under model m_j .

There are many other Bayesian criteria for model selection. The advantage of criterion based methods for model selection is that:

1. one does not need to specify a prior on the model space
2. improper priors can be used to define the criteria.

On the other hand, posterior model probabilities require:

1. proper prior distributions for all of the parameters arising from the k possible models in the model space. There are many priors to elicit (k proper priors).
2. a prior distribution on the model space \mathcal{M} .

Thus computing posterior model probabilities can be quite difficult in practice, especially when k is large. For example, in variable subset selection with 10 covariates, $k = 2^{10} = 1024$. Here, we would have to specify prior distributions for the regression coefficients for 1024 models, and we would need to specify 1024 probabilities for the model space. When $k = 30$, we would need to specify $2^{30} \approx 10^9$ probabilities for the model space.

Criterion based methods do not have such a requirement. Most criterion based methods only require prior distributions on the parameters arising from the different models. BIC does not even require priors. However, criterion based methods, even with a Bayesian motivation, are not fully Bayesian.

Example 1.27: Variable Selection in Linear Regression

Consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where the ϵ_i 's are i.i.d. and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$.

We are interested in variable subset selection. In this problem, \mathcal{M} consists of $k = 4$ models. In variable subset selection problems, we always include the intercept in every model by convention.

Thus, we have

m_1 : (intercept model)

m_2 : (x_1)

m_3 : (x_2)

m_4 : (x_1, x_2)

The notation (x_1) means the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i,$$

and so forth.

The regression coefficients have a different physical meaning from model to model, and thus to specify prior distributions, we have to think about the β 's differently for the 4 models.

Prior for the model space

We have

$$\mathcal{M} = \{m_1, \dots, m_4\} .$$

A uniform prior on \mathcal{M} is $p(m_1) = \dots = p(m_4) = \frac{1}{4}$.

Another prior might be $p(m_1) = .1$, $p(m_2) = .3$, $p(m_3) = .4$, $p(m_4) = .2$, and so on. We need good methods to specify a prior for \mathcal{M} because a uniform prior for \mathcal{M} may not be desirable if k is large, say $k = 1024$.

Prior Distributions for the Regression Parameters

We have

m_1 :	Intercept	(β_0)
m_2 :	(x_1)	(β_0, β_1)
m_3 :	(x_2)	(β_0, β_2)
m_4 :	(x_1, x_2)	$(\beta_0, \beta_1, \beta_2)$.

$\tau = \frac{1}{\sigma^2}$ also needs a prior distribution. Thus we need to specify prior distributions for 9 parameters. Model selection for the linear model will be discussed in much more detail shortly.

Highest Posterior Density Regions

The Bayesian “confidence interval” is called a highest posterior density (HPD) region (or a credible set). HPD regions and credible sets are *not* the same thing in general.

Suppose θ is a univariate parameter. A 95% HPD interval for θ is the interval such that 95% of the highest area of the posterior density is contained in this interval.

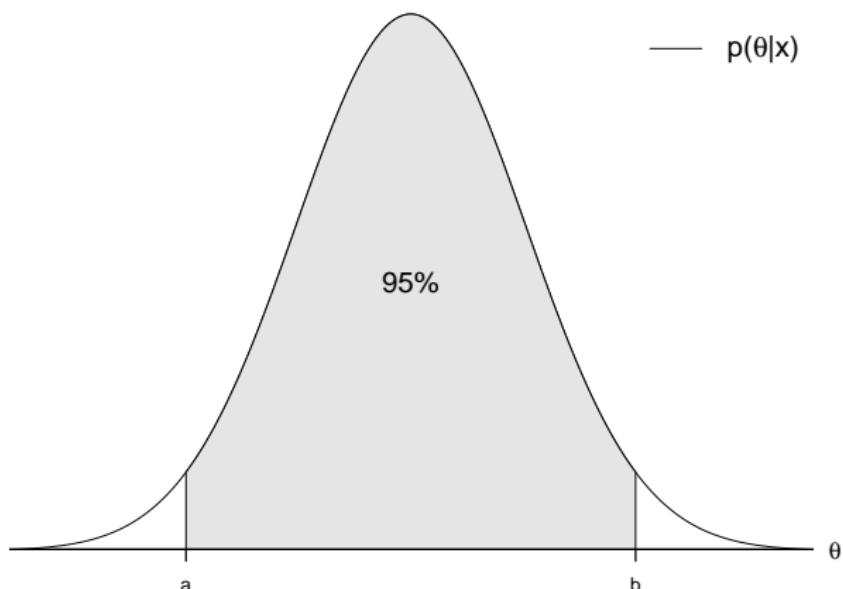


Figure 1.12:

A 95% HPD interval for θ is $\{\theta : a \leq \theta \leq b\}$.

If θ is multidimensional, they are called **HPD regions**. A formal definition is given as follows:

Definition 1.4

Let $p(\theta|x)$ denote the posterior density of θ . A region R in the parameter space Θ is called an HPD region of content $1 - \alpha$ if

- (a) $P(\theta \in R|x) = 1 - \alpha$
- (b) for $\theta_1 \in R$ and $\theta_2 \notin R$, $p(\theta_1|x) \geq p(\theta_2|x)$.

Remark 1.11

A credible set is a posterior region that is constructed by removing the upper $1 - \alpha/2$ and lower $1 - \alpha/2$ percentiles of the posterior distribution. Such a set is called a $(1 - \alpha) \times 100\%$ credible set.

HPD region = credible set when the posterior distribution is symmetric.

When a posterior distribution is skewed, it may be hard to construct HPD regions. Credible sets are easier to construct in this case. When the distribution is symmetric, HPD regions are easy to construct. Often, the credible set \approx HPD region, especially for large samples.

Some Properties of HPD regions

1. It follows from the definition that for a given probability content $1 - \alpha$, the HPD region has the smallest possible volume in the parameter space Θ .
2. If we make the assumption that $p(\theta|x)$ is non-uniform over every region in the space of θ , then the HPD region of content $1 - \alpha$ is unique. Further if θ_1 and θ_2 are two points such that $p(\theta_1|x) = p(\theta_2|x)$, then these two points are simultaneously included (or excluded) by a $1 - \alpha$ HPD region. The converse is also true.

That is, if $p(\theta_1|x) \neq p(\theta_2|x)$, then there exists a $1 - \alpha$ HPD region which includes one point but not the other.

Example 1.28

Suppose x_1, \dots, x_n are i.i.d. $N(\theta, 1)$, and take $\pi(\theta) \propto 1$. Let $x = (x_1, \dots, x_n)$. We have established that

$$\begin{aligned}\theta|x &\sim N\left(\bar{x}, \frac{1}{n}\right), \\ z|x &\sim N\left(\bar{x}, 1 + \frac{1}{n}\right).\end{aligned}$$

Note that the posterior mean of θ is the frequentist point estimate of θ , which is \bar{x} . The posterior variance of θ is the frequentist variance of \bar{x} , i.e., $\text{Var}(\bar{x}|\theta) = \frac{1}{n}$. Also, the predictive mean of z is the frequentist point estimate of a future value, and the predictive variance of z is the frequentist estimate. That is,

$$\text{Var}(z - \bar{x}|\theta) = 1 + \frac{1}{n}.$$

Now let us construct a 95% HPD interval for θ . We know that $\theta|x \sim N(\bar{x}, \frac{1}{n})$.

We have $P(\bar{x} - 1.96 \frac{1}{\sqrt{n}} \leq \theta \leq \bar{x} + 1.96 \frac{1}{\sqrt{n}} | x) = .95$. Thus a 95% HPD interval for θ is

$$\left\{ \theta : \bar{x} - \frac{1.96}{\sqrt{n}} \leq \theta \leq \bar{x} + \frac{1.96}{\sqrt{n}} \right\}.$$

Notice that this is the same interval as a frequentist confidence interval for θ , but it has a much different interpretation.

Suppose we want to construct a 95% **Highest Predictive Density** interval for z . This interval is given by

$$\left\{ z : \bar{x} - 1.96 \sqrt{1 + \frac{1}{n}} \leq z \leq \bar{x} + 1.96 \sqrt{1 + \frac{1}{n}} \right\}.$$

This is the same interval as the frequentist predictive interval for z .

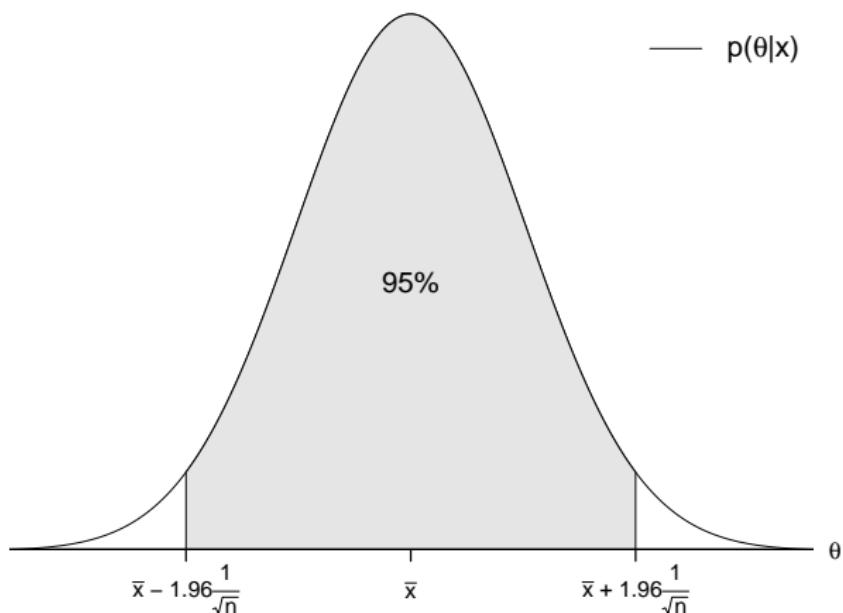


Figure 1.13:

Example 1.29

Suppose x_1, \dots, x_n are i.i.d. $N(\mu, \sigma^2)$, where (μ, σ^2) are both unknown. Let $\tau = \frac{1}{\sigma^2}$. Suppose we specify the joint prior

$$\pi(\mu, \tau) \propto \tau^{-1}.$$

Now,

$$p(\mu, \tau | x) \propto \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} \sum (x_i - \mu)^2 \right\}.$$

Thus

$$\begin{aligned} p(\mu | x) &\propto \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} \sum (x_i - \mu)^2 \right\} d\tau \\ &= \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} [\Sigma x_i^2 - 2\mu \Sigma x_i + n\mu^2] \right\} d\tau \\ &= \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ \frac{n\tau}{2} \bar{x}^2 \right\} \exp \left\{ -\frac{n\tau}{2} \left[\frac{1}{n} \Sigma x_i^2 \right] \right\} \exp \left\{ -\frac{n\tau}{2} [\mu - \bar{x}]^2 \right\} d\tau \\ &= \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} [\Sigma x_i^2 - n\bar{x}^2] \right\} \exp \left\{ -\frac{n\tau}{2} [\mu - \bar{x}]^2 \right\} d\tau \\ &= \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} [\Sigma (x_i - \bar{x})^2] \right\} \exp \left\{ -\frac{n\tau}{2} [\mu - \bar{x}]^2 \right\} d\tau. \end{aligned}$$

Let $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$. Now, we get

$$\begin{aligned} &= \int_0^\infty \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [(n-1)s^2 + n(\mu - \bar{x})^2]\right\} d\tau \\ &\propto [(n-1)s^2 + n(\mu - \bar{x})^2]^{-\frac{n}{2}} \\ &\propto \left[1 + \frac{n}{(n-1)s^2} (\mu - \bar{x})^2\right]^{-\frac{(n-1+1)}{2}}. \end{aligned}$$

Thus

$$\begin{aligned} \mu|x &\sim S_1(n-1, \bar{x}, \frac{s^2}{n}) \\ &= t(n-1, \bar{x}, \frac{s^2}{n}). \end{aligned}$$

Hence $\frac{\mu - \bar{x}}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$.

Therefore, a 95% HPD interval for μ is

$$\left\{ \mu : \bar{x} - t_{(n-1,.975)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1,.975)} \frac{s}{\sqrt{n}} \right\},$$

where $t_{(n-1,.975)}$ corresponds to the 97.5th percentile of the t distribution with $n - 1$ degrees of freedom. This HPD interval corresponds to the frequentist confidence interval for μ when σ^2 is unknown.

Let us construct a credible interval for $\tau = \frac{1}{\sigma^2}$.

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

Thus $\tau|x \sim \text{gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$.

Note that the above posterior distribution for τ implies

$$(n - 1)s^2 \tau \sim \chi_{n-1}^2.$$

Thus a 95% credible interval for τ is given by

$$\chi_{(n-1,.025)}^2 \leq (n - 1)s^2 \tau \leq \chi_{(n-1,.975)}^2$$

and therefore,

$$\left\{ \tau : \frac{\chi_{(n-1,.025)}^2}{(n - 1)s^2} \leq \tau \leq \frac{\chi_{(n-1,.975)}^2}{(n - 1)s^2} \right\}.$$

This is the same interval as the 95% frequentist confidence interval for $\tau = \frac{1}{\sigma^2}$. Let us now find the predictive distribution of a future observation z .

$$\begin{aligned}
p(z|x) &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2}\Sigma(x_i - \mu)^2\right\} \tau^{\frac{1}{2}} \exp\left\{-\frac{\tau}{2}(z - \mu)^2\right\} d\mu d\tau \\
&= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+1}{2}-1} \exp\left\{-\frac{\tau}{2} [\Sigma(x_i - \mu)^2 + (z - \mu)^2]\right\} d\mu d\tau \\
&= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+1}{2}-1} \exp\left\{-\frac{\tau}{2} [(n-1)s^2 + n(\mu - \bar{x})^2 + (z - \mu)^2]\right\} d\mu d\tau \\
&= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+1}{2}-1} \exp\left\{-\frac{\tau}{2} [(n-1)s^2 + \mu^2(1+n) - 2\mu(n\bar{x} + z)\right. \\
&\quad \left.+ (z^2 + n\bar{x}^2)]\right\} d\mu d\tau \\
&= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+1}{2}-1} \exp\left\{-\frac{\tau}{2} [z^2 + n\bar{x}^2]\right\} \exp\left\{-\frac{\tau}{2} [(n-1)s^2]\right\} \\
&\quad \times \exp\left\{-\frac{(n+1)\tau}{2} \left[\mu - \frac{(n\bar{x} + z)}{n+1}\right]^2\right\} \exp\left\{\frac{(n+1)\tau}{2} \left(\frac{n\bar{x} + z}{n+1}\right)^2\right\} d\mu d\tau
\end{aligned}$$

$$\begin{aligned}
& \propto \int_0^\infty \tau^{\frac{n+1}{2}-1} \exp \left\{ -\frac{\tau}{2} [z^2 + n\bar{x}^2] \right\} \exp \left\{ -\frac{\tau}{2} [(n-1)s^2] \right\} \exp \left\{ \frac{\tau}{2} \frac{[n\bar{x} + z]^2}{n+1} \right\} \tau^{-\frac{1}{2}} d\tau \\
& = \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[z^2 \left(\frac{n}{n+1} \right) + n\bar{x}^2 \left(\frac{1}{n+1} \right) + (n-1)s^2 - \frac{2zn\bar{x}}{n+1} \right] \right\} d\tau \\
& = \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[\frac{n\bar{x}^2}{n+1} + (n-1)s^2 \right] \right\} \exp \left\{ -\frac{n\tau}{2(n+1)} [z^2 - 2z\bar{x}] \right\} d\tau \\
& = \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[\frac{n\bar{x}^2}{n+1} + (n-1)s^2 \right] \right\} \exp \left\{ -\frac{n\tau}{2(n+1)} [z - \bar{x}]^2 \right\} \exp \left\{ \frac{n\tau\bar{x}^2}{2(n+1)} \right\} d\tau
\end{aligned}$$

$$\begin{aligned}
&= \int_0^\infty \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[(n-1)s^2 + \frac{n}{n+1}(z-\bar{x})^2 \right] \right\} d\tau \\
&\propto \left[(n-1)s^2 + \frac{n}{n+1}(z-\bar{x})^2 \right]^{-\frac{n}{2}} \\
&\propto \left[1 + \left(\frac{n}{(n+1)s^2} \right) \left(\frac{1}{n-1} \right) (z-\bar{x})^2 \right]^{-\frac{(n-1+1)}{2}}.
\end{aligned}$$

Thus

$$z|x \sim t(n-1, \bar{x}, s^2(1 + \frac{1}{n})) .$$

The mean of the predictive distribution is the same as the frequentist point estimate of z , and note that $s^2(1 + \frac{1}{n})$ is the frequentist estimate of $\text{Var}(z - \bar{x}|\mu, \sigma^2)$.

Note that

$$\begin{aligned}\text{Var}(z - \bar{x}|\mu, \sigma^2) &= \text{Var}(z|\mu, \sigma^2) + \text{Var}(\bar{x}|\mu, \sigma^2) - 2\text{Cov}(z, \bar{x}|\mu, \sigma^2) \\ &= \sigma^2 + \frac{\sigma^2}{n} - 0 \\ &= \sigma^2\left(1 + \frac{1}{n}\right).\end{aligned}$$

Thus $\text{Var}(z - \bar{x}|\hat{\mu}, \hat{\sigma}^2) = s^2\left(1 + \frac{1}{n}\right)$.

A 95% highest predictive density interval is given by

$$\left\{ z : \bar{x} - t_{(n-1,.975)}s\sqrt{1 + \frac{1}{n}} \leq z \leq \bar{x} + t_{(n-1,.975)}s\sqrt{1 + \frac{1}{n}} \right\},$$

which is the same as the frequentist prediction interval for z .

Example 1.30

The R function below calculates the $(1 - \alpha) \times 100\%$ HPD interval for any previously derived $\text{beta}(\beta, \lambda)$ posterior distribution.

Consider Example 1.23 from page 110. We derive the 95% HPD interval for the $\text{beta}(23, 8)$ posterior distribution:

```
hpd.beta <- function(alpha, beta, lambda){  
  q.left <- 0  
  q.right <- 1  
  diff <- q.right - q.left  
  alpha.temp <- 0.001  
  while(alpha.temp <= alpha)  
  {  
    q.left.temp <- qbeta(alpha.temp, beta, lambda)  
    q.right.temp <- qbeta(.95 + alpha.temp, beta, lambda)  
    diff.temp <- round(q.right.temp - q.left.temp, 5)  
    if(diff.temp >= diff) break  
    q.left <- q.left.temp  
    q.right <- q.right.temp  
    diff <- diff.temp  
    alpha.temp <- alpha.temp + 0.001  
  }  
  return( c(q.left, q.right) )  
}
```

```
# 95% HPD interval for the beta(23,8) posterior distribution  
> hpd.beta(.05,23,8)  
[1] 0.5883438 0.8854985
```

Chapter 2:

Bayesian Methods for the Linear Model

Bayesian Analysis of the Linear Model

The linear model is frequently used in many biostatistical applications, including

1. dose response modeling
2. polynomial regression
3. exposure assessment
4. analysis of variance problems for comparing treatment groups

(See *Case Studies in Biometry* by Lange et al., John Wiley & Sons.)

The linear model can be written as

$$Y = X\beta + \epsilon , \tag{2.1}$$

where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$ and

$$\epsilon \sim N_n(0, \sigma^2 I) . \tag{2.2}$$

Let $M = X(X'X)^{-}X'$, and $\tau = \frac{1}{\sigma^2}$, where the $-$ denotes generalized inverse. Recall that the UMVUE of $\mu = E(Y) = X\beta$ is MY .

We would like to derive the posterior distributions of β and τ under noninformative priors.

Theorem 2.1

Suppose τ is known, X is of full rank p , and

$$\pi(\beta) \propto 1.$$

Then

$$\beta|y, \tau \sim N_p(\hat{\beta}, \tau^{-1}(X'X)^{-1}) ,$$

where

$$\hat{\beta} = (X'X)^{-1}X'Y .$$

Proof:

$$p(\beta|y, \tau) \propto \exp \left\{ -\frac{\tau}{2} (Y - X\beta)'(Y - X\beta) \right\}$$

Note that

$$\begin{aligned} & Y'(I - M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ &= Y'(I - M)Y + \beta'X'X\beta - 2\hat{\beta}'X'X\beta + \hat{\beta}'X'X\hat{\beta} \\ &= Y'(I - M)Y + \beta'X'X\beta - 2Y'X(X'X)^{-1}(X'X)\beta + Y'MY \\ &= Y'Y + \beta'X'X\beta - 2Y'X\beta \\ &= (Y - X\beta)'(Y - X\beta). \end{aligned}$$

Thus

$$\begin{aligned} p(\beta|y, \tau) &\propto \exp \left\{ -\frac{\tau}{2} (Y - X\beta)'(Y - X\beta) \right\} \\ &= \exp \left\{ -\frac{\tau}{2} \left[Y'(I - M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \right] \right\} \\ &\propto \exp \left\{ -\frac{\tau}{2} \left[(\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \right] \right\}. \end{aligned}$$

We can recognize this as a normal kernel with mean $\hat{\beta}$ and covariance matrix $\tau^{-1}(X'X)^{-1}$. Thus

$$\beta|y, \tau \sim N_p(\hat{\beta}, \tau^{-1}(X'X)^{-1}).$$

The posterior and predictive distributional results for the linear model are a generalization of the i.i.d. case discussed on previous pages.

To obtain the i.i.d. case for the linear model, we set $X = 1_{n \times 1}$ where $1_{n \times 1}$ is a $n \times 1$ vector of ones, and we write

$$y_{n \times 1} = 1_{n \times 1}\beta_{1 \times 1} + \epsilon_{n \times 1}, \quad \epsilon \sim N(0, \sigma^2),$$

which implies that the y_i 's are i.i.d. $N(\beta, \sigma^2)$, $i = 1, \dots, n$.

Substituting $X = 1_{n \times 1}$, for example, in the previous theorem, we get

$$\beta | y, \tau \sim N_1(\bar{y}, \tau^{-1}n^{-1}),$$

where

$$(X'X)^{-1}X'y = n^{-1}1'y = \bar{y}, \quad \text{and} \quad (X'X)^{-1} = (1'1)^{-1} = 1/n.$$

Theorem 2.2

When τ is known, Jeffreys prior for β is a uniform prior, i.e.,

$$\pi(\beta) \propto 1.$$

Proof:

$$\begin{aligned}\log[p(y|\beta, \tau)] &= -\frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\tau) - \frac{\tau}{2} (Y - X\beta)'(Y - X\beta). \\ \frac{\partial}{\partial \beta} \log[p(y|\beta, \tau)] &= \frac{\partial}{\partial \beta} \left[-\frac{\tau}{2} (Y - X\beta)'(Y - X\beta) \right] \\ &= \frac{\partial}{\partial \beta} \left[-\frac{\tau}{2} [Y'Y - 2\beta'X'Y + \beta'X'X\beta] \right] \\ &= \tau X'Y - \tau(X'X)\beta.\end{aligned}$$

Also,

$$\frac{\partial^2}{\partial \beta \partial \beta'} \log[p(y|\beta, \tau)] = -\tau(X'X),$$

and therefore,

$$I(\beta) = \tau(X'X).$$

Thus Jeffreys prior for β is given by

$$\pi(\beta|\tau) \propto |\tau(X'X)|^{\frac{1}{2}} \propto \text{constant}.$$

Thus

$$\pi(\beta|\tau) \propto 1.$$

Theorem 2.3

Consider the linear model in (2.1) and (2.2) with both β and τ unknown. Then Jeffreys joint prior for (β, τ) is given by

$$\pi(\beta, \tau) \propto \tau^{\frac{p}{2}-1} .$$

Proof:

$$\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \beta \partial \beta'} = -\tau(X'X) .$$

$$\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \tau^2} = -\frac{n}{2\tau^2} .$$

$$\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \beta \partial \tau} = X'Y - (X'X)\beta .$$

Note that

$$\begin{aligned} -E\left(\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \beta \partial \tau}\right) &= -E(X'Y) + (X'X)\beta \\ &= -(X'X)\beta + (X'X)\beta \\ &= 0 . \end{aligned}$$

Thus,

$$I(\beta, \tau) = \begin{bmatrix} \tau(X'X) & 0 \\ 0 & \frac{n}{2\tau^2} \end{bmatrix}.$$

Now

$$\begin{aligned} |I(\beta, \tau)| &= |\tau(X'X)| \left| \frac{n}{2} \tau^{-2} \right| \\ &= \tau^p |X'X| \frac{n}{2} \tau^{-2} \\ &= \tau^{p-2} \frac{n}{2} |X'X| \\ &\propto \tau^{p-2}. \end{aligned}$$

Thus

$$\begin{aligned} \pi(\beta, \tau) &\propto |I(\beta, \tau)|^{\frac{1}{2}} \\ &= \tau^{\frac{p-2}{2}} = \tau^{\frac{p}{2}-1}. \end{aligned}$$

Notice that with $p = 1$ (i.i.d. case),

$$\pi(\beta, \tau) \propto \tau^{-\frac{1}{2}}.$$

Theorem 2.4

Consider the linear model in (2.1) and (2.2) with (β, τ) unknown, and suppose

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Then

$$\beta|y \sim S_p(n-p, \hat{\beta}, s^2(X'X)^{-1}),$$

where $s^2 = \frac{Y'(I-M)Y}{n-p}$ and

$$\tau|y \sim \text{gamma}\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right).$$

Proof:

We have

$$\begin{aligned} p(\beta, \tau|y) &\propto \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [(Y - X\beta)'(Y - X\beta)]\right\} \\ &= \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I - M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]\right\}. \end{aligned}$$

Thus

$$\begin{aligned} p(\beta|y) &\propto \int_0^{\infty} \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I-M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]\right\} d\tau \\ &\propto [Y'(I-M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]^{-\frac{n}{2}} \end{aligned}$$

Let $s^2 = \frac{Y'(I-M)Y}{n-p}$. Then the above integral is

$$\begin{aligned} &= [(n-p)s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]^{-\frac{(n-p+p)}{2}} \\ &\propto \left[1 + \frac{1}{s^2(n-p)}(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]^{-\frac{(n-p+p)}{2}}. \end{aligned}$$

Thus

$$\beta|y \sim S_p \left(n-p, \hat{\beta}, s^2(X'X)^{-1} \right).$$

Now

$$\begin{aligned}
 p(\tau|y) &\propto \int_{-\infty}^{\infty} \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I-M)Y + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})]\right\} d\beta \\
 &= \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I-M)Y]\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{\tau}{2}(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})\right\} d\beta \\
 &\propto \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I-M)Y]\right\} \tau^{-\frac{p}{2}} \\
 &= \tau^{\frac{n-p}{2}-1} \exp\left\{-\frac{\tau}{2} [(n-p)s^2]\right\} .
 \end{aligned}$$

Thus

$$\tau|y \sim \text{gamma}\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right).$$

Theorem 2.5

Consider the linear model in (2.1) and (2.2) with (β, τ) unknown, and suppose

$$\begin{aligned}\beta | \tau &\sim N_p(\mu_0, \tau^{-1} \Sigma_0), \\ \tau &\sim \text{gamma } \left(\frac{\delta_0}{2}, \frac{\gamma_0}{2} \right).\end{aligned}$$

Then

$$\beta | y \sim S_p \left(n + \delta_0, \tilde{\beta}, \tilde{s}^2 (X'X + \Sigma_0^{-1})^{-1} \right),$$

where

$$\begin{aligned}\tilde{\beta} &= \Lambda \mu_0 + (I - \Lambda) \hat{\beta}, \\ \Lambda &= (X'X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1}, \\ \hat{\beta} &= (X'X)^{-1} X'Y,\end{aligned}$$

$$\tilde{s}^2 = (n + \delta_0)^{-1} \left[Y'(I - M)Y + (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0) + \gamma_0 \right],$$

and

$$\tau | y \sim \text{gamma} \left(\frac{n + \delta_0}{2}, \frac{(n + \delta_0)\tilde{s}^2}{2} \right).$$

Proof:

$$p(\beta|y) \propto \int_0^{\infty} \tau^{\frac{n+p+\delta_0}{2}-1} \times \exp\left\{-\frac{\tau}{2} [\gamma_0 + Y'(I-M)Y + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + (\beta - \mu_0)' \Sigma_0^{-1}(\beta - \mu_0)]\right\} d\tau .$$

Now

$$\begin{aligned} & (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + (\beta - \mu_0)' \Sigma_0^{-1}(\beta - \mu_0) \\ &= (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) - \tilde{\beta}'(X'X + \Sigma_0^{-1})\tilde{\beta} + \hat{\beta}'X'X\hat{\beta} + \mu_0'\Sigma_0^{-1}\mu_0 \end{aligned}$$

Note that

$$\begin{aligned} \Lambda &= (X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}, \\ I - \Lambda &= (X'X + \Sigma_0^{-1})^{-1}X'X, \end{aligned}$$

since

$$\begin{aligned} I - \Lambda + \Lambda &= (X'X + \Sigma_0^{-1})^{-1}X'X + (X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1} \\ &= (X'X + \Sigma_0^{-1})^{-1}(X'X + \Sigma_0^{-1}) = I. \end{aligned}$$

Now observe that

$$\begin{aligned}
 & -\tilde{\beta}'(X'X + \Sigma_0^{-1})\tilde{\beta} + \hat{\beta}'X'X\hat{\beta} + \mu_0'\Sigma_0^{-1}\mu_0 \\
 &= -(\Lambda\mu_0 + (I - \Lambda)\hat{\beta})'(X'X + \Sigma_0^{-1})(\Lambda\mu_0 + (I - \Lambda)\hat{\beta}) + Y'MY + \mu_0'\Sigma_0^{-1}\mu_0 \\
 &= -\mu_0'\Lambda' \left(X'X + \Sigma_0^{-1} \right) \Lambda\mu_0 - \mu_0'\Lambda' \left(X'X + \Sigma_0^{-1} \right) (I - \Lambda)\hat{\beta} \\
 &\quad - \hat{\beta}'(I - \Lambda)' \left(X'X + \Sigma_0^{-1} \right) \Lambda\mu_0 \\
 &\quad - \hat{\beta}'(I - \Lambda)' \left(X'X + \Sigma_0^{-1} \right) (I - \Lambda)\hat{\beta} + Y'MY + \mu_0'\Sigma_0^{-1}\mu_0 .
 \end{aligned}$$

Now

$$\begin{aligned}
 & -\mu_0'\Lambda' \left(X'X + \Sigma_0^{-1} \right) \Lambda\mu_0 + \mu_0'\Sigma_0^{-1}\mu_0 \\
 &= -\mu_0' \left(\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1} \right) (X'X + \Sigma_0^{-1})(X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}\mu_0 + \mu_0'\Sigma_0^{-1}\mu_0 \\
 &= -\mu_0'\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}\mu_0 + \mu_0'\Sigma_0^{-1}\mu_0 \\
 &= \mu_0'\Sigma_0^{-1} \left(I - (X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1} \right) \mu_0 \\
 &= \mu_0'\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}X'X\mu_0 \\
 &= \mu_0'\Lambda'X'X\mu_0 .
 \end{aligned}$$

Now

$$\begin{aligned}
 & -2\mu_0' \Lambda' \left(X'X + \Sigma_0^{-1} \right) (I - \Lambda) \hat{\beta} \\
 &= -2\mu_0' \Sigma_0^{-1} (X'X + \Sigma_0^{-1})^{-1} (X'X + \Sigma_0^{-1}) (X'X + \Sigma_0^{-1})^{-1} X'X \hat{\beta} \\
 &= -2\mu_0' \Sigma_0^{-1} (X'X + \Sigma_0^{-1})^{-1} X'X \hat{\beta} \\
 &= -2\mu_0' \Lambda' X'X \hat{\beta}.
 \end{aligned}$$

Finally

$$\begin{aligned}
 & -\hat{\beta}' (I - \Lambda)' (X'X + \Sigma_0^{-1}) (I - \Lambda) \hat{\beta} \\
 &= -\hat{\beta}' \left(X'X (X'X + \Sigma_0^{-1})^{-1} (X'X + \Sigma_0^{-1}) (X'X + \Sigma_0^{-1})^{-1} X'X \right) \hat{\beta} \\
 &= -\hat{\beta}' X'X (X'X + \Sigma_0^{-1})^{-1} X'X \hat{\beta} \\
 &= -\hat{\beta}' (I - \Lambda') X'X \hat{\beta} \\
 &= -\hat{\beta}' X'X \hat{\beta} + \hat{\beta}' \Lambda' X'X \hat{\beta} \\
 &= -Y'MY + \hat{\beta}' \Lambda' X'X \hat{\beta}.
 \end{aligned}$$

Thus

$$\begin{aligned}
 & -\tilde{\beta}' (X'X + \Sigma_0^{-1}) \tilde{\beta} + \hat{\beta}' X'X \hat{\beta} + \mu_0' \Sigma_0^{-1} \mu_0 \\
 &= \mu_0' \Lambda' X'X \mu_0 - 2\mu_0' \Lambda' X'X \hat{\beta} - Y'MY + \hat{\beta}' \Lambda' X'X \hat{\beta} + Y'MY \\
 &= \mu_0' \Lambda' X'X \mu_0 - 2\mu_0' \Lambda' X'X \hat{\beta} + \hat{\beta}' \Lambda' X'X \hat{\beta} \\
 &= (\hat{\beta} - \mu_0)' (\Lambda' X'X) (\hat{\beta} - \mu_0).
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(\beta|y) & \propto \int_0^\infty \tau^{\frac{n+p+\delta_0}{2}-1} \\
 & \times \exp \left\{ -\frac{\tau}{2} \left[\gamma_0 + Y'(I-M)Y + (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) + (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0) \right] \right\} d\tau \\
 & = \int_0^\infty \tau^{\frac{n+p+\delta_0}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[(n + \delta_0)\tilde{s}^2 + (\beta - \tilde{\beta})' \left(X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}) \right] \right\} d\tau \\
 & \propto \left[(n + \delta_0)\tilde{s}^2 + (\beta - \tilde{\beta})' \left(X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}) \right]^{-\frac{(n+\delta_0+p)}{2}} \\
 & \propto \left[1 + \frac{1}{(n + \delta_0)\tilde{s}^2} (\beta - \tilde{\beta})' \left(X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}) \right]^{-\frac{(n+\delta_0+p)}{2}}.
 \end{aligned}$$

Thus

$$\beta|y \sim S_p \left(n + \delta_0, \tilde{\beta}, \tilde{s}^2(X'X + \Sigma_0^{-1})^{-1} \right).$$

Now

$$\begin{aligned}
 p(\tau|y) &\propto \int_{-\infty}^{\infty} \tau^{\frac{n+p+\delta_0}{2}-1} \\
 &\quad \times \exp\left\{-\frac{\tau}{2} \left[(n+\delta_0)\tilde{s}^2 + (\beta - \tilde{\beta})' \left(X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}) \right] \right\} d\beta \\
 &= \tau^{\frac{n+p+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2} [(n+\delta_0)\tilde{s}^2]\right\} \\
 &\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{\tau}{2} \left[(\beta - \tilde{\beta})' \left(X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}) \right] \right\} d\beta \\
 &\propto \tau^{\frac{n+p+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2} [(n+\delta_0)\tilde{s}^2]\right\} \tau^{-\frac{p}{2}} \\
 &= \tau^{\frac{n+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2} [(n+\delta_0)\tilde{s}^2]\right\}.
 \end{aligned}$$

Thus

$$\tau|y \sim \text{gamma}\left(\frac{n+\delta_0}{2}, \frac{(n+\delta_0)\tilde{s}^2}{2}\right).$$

Theorem 2.6

Consider the linear model in (2.1) and (2.2). Let Z be a $q \times 1$ vector of future observations taken at X_f , where X_f is $q \times p$. That is

$$Z = X_f\beta + \epsilon,$$

where

$$\epsilon \sim N_q(0, \sigma^2 I).$$

Suppose

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Then

$$Z|X_f, Y \sim S_q \left(n - p, X_f \hat{\beta}, s^2 (I + X_f (X'X)^{-1} X_f') \right),$$

where

$$s^2 = \frac{Y'(I - M)Y}{n - p}.$$

Proof:

$$\begin{aligned}
 p(z|X_f, y) &= \int_0^\infty \int_{-\infty}^\infty p(z|\beta, \tau) p(\beta, \tau|y) d\beta d\tau \\
 &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{q}{2}} \exp\left\{-\frac{\tau}{2} [(z - X_f\beta)'(z - X_f\beta)]\right\} \tau^{\frac{n}{2}-1} \\
 &\quad \times \exp\left\{-\frac{\tau}{2} \left[(n-p)s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\right\} d\beta d\tau .
 \end{aligned}$$

Note that

$$(z - X_f\beta)'(z - X_f\beta) = (\beta - \hat{\beta}_z)'(X_f'X_f)(\beta - \hat{\beta}_z) + z'(I - M_{x_f})z ,$$

where

$$\hat{\beta}_z = (X_f'X_f)^{-1}X_f'z ,$$

$$M_{x_f} = X_f(X_f'X_f)^{-1}X_f' .$$

Now

$$\begin{aligned}
 & (\beta - \hat{\beta}_z)'(X_f'X_f)(\beta - \hat{\beta}_z) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\
 = & (\beta - \tilde{\beta}_f)'(X_f'X_f + X'X)(\beta - \tilde{\beta}_f) \\
 & - \tilde{\beta}_f'(X_f'X_f + X'X)\tilde{\beta}_f + \hat{\beta}_z'X_f'X_f\hat{\beta}_z + \hat{\beta}'X'X\hat{\beta} ,
 \end{aligned}$$

where

$$\tilde{\beta}_f = (X_f'X_f + X'X)^{-1} \left[(X_f'X_f)\hat{\beta}_z + (X'X)\hat{\beta} \right]$$

$$= \Lambda_f \hat{\beta}_z + (I - \Lambda_f) \hat{\beta} ,$$

$$\Lambda_f = (X_f'X_f + X'X)^{-1}X_f'X_f .$$

Note that

$$I - \Lambda_f = (X'_f X_f + X' X)^{-1} X' X$$

since

$$\begin{aligned} I - \Lambda_f + \Lambda_f &= (X'_f X_f + X' X)^{-1} X' X + (X'_f X_f + X' X)^{-1} X'_f X_f \\ &= (X'_f X_f + X' X)^{-1} (X' X + X'_f X_f) = I. \end{aligned}$$

Now

$$\begin{aligned} &- \tilde{\beta}'_f (X'_f X_f + X' X) \tilde{\beta}_f + \hat{\beta}'_z X'_f X_f \hat{\beta}_z + \hat{\beta}' X' X \hat{\beta} \\ &= (\hat{\beta} - \hat{\beta}_z)' (\Lambda'_f X' X) (\hat{\beta} - \hat{\beta}_z), \end{aligned}$$

as in the previous proof.

Thus

$$\begin{aligned} p(z|X_f, y) &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+q}{2}-1} \exp \left\{ -\frac{\tau}{2} \left[(n-p)s^2 + (\beta - \tilde{\beta}_f)' (X'_f X_f + X' X)^{-1} (\beta - \tilde{\beta}_f) \right] \right\} \\ &\quad \times \exp \left\{ -\frac{\tau}{2} \left[(\hat{\beta} - \hat{\beta}_z)' (\Lambda'_f X' X) (\hat{\beta} - \hat{\beta}_z) + z' (I - M_{X_f}) z \right] \right\} d\beta d\tau. \\ &= \int_0^\infty \tau^{\frac{n+q}{2}-1} \tau^{-\frac{p}{2}} \\ &\quad \times \exp \left\{ -\frac{\tau}{2} \left[(n-p)s^2 + (\hat{\beta} - \hat{\beta}_z)' (\Lambda'_f X' X) (\hat{\beta} - \hat{\beta}_z) + z' (I - M_{X_f}) z \right] \right\} d\tau. \end{aligned}$$

Now write

$$\begin{aligned}
 & (\hat{\beta} - \hat{\beta}_z)'(\Lambda_f' X' X)(\hat{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z \\
 &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2\hat{\beta}_z' \Lambda_f' X' X \hat{\beta} + \hat{\beta}_z' \Lambda_f' X' X \hat{\beta}_z + z'(I - M_{x_f})z \\
 &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f)^{-1} (X_f' X_f) (X_f' X_f + X' X)^{-1} (X' X) \hat{\beta} \\
 &\quad + z' X_f (X_f' X_f)^{-1} (X_f' X_f) (X_f' X_f + X' X)^{-1} (X' X) (X_f' X_f)^{-1} X_f' z \\
 &\quad + z'(I - M_{x_f})z \\
 &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\
 &\quad + z' X_f (I - \Lambda_f) (X_f' X_f)^{-1} X_f' z + z'(I - M_{x_f})z \\
 &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\
 &\quad + z' X_f (X_f' X_f)^{-1} X_f' z - z' X_f \Lambda_f (X_f' X_f)^{-1} X_f' z \\
 &\quad + z' z - z' X_f (X_f' X_f)^{-1} X_f' z \\
 &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\
 &\quad + z'(I - X_f (X_f' X_f + X' X)^{-1} X_f') z .
 \end{aligned}$$

Note that

$$\begin{aligned}
 \Lambda_f (X_f' X_f)^{-1} &= (X_f' X_f + X' X)^{-1} (X_f' X_f) (X_f' X_f)^{-1} \\
 &= (X_f' X_f + X' X)^{-1} .
 \end{aligned}$$

Now we need to establish two non-trivial identities in order to finish completing the square.

Identity 2.1

$$(I - X_f(X'_f X_f + X'X)^{-1} X'_f)^{-1} = I + X_f(X'X)^{-1} X'_f .$$

Proof:

It suffices to show that

$$[I - X_f(X'X + X'_f X_f)^{-1} X'_f] [I + X_f(X'X)^{-1} X'_f] = I .$$

Let's multiply the left and show the right. Doing this, we get

$$\begin{aligned} & I - X_f(X'X + X'_f X_f)^{-1} X'_f - X_f(X'X + X'_f X_f)^{-1} X'_f X_f(X'X)^{-1} X'_f \\ & \quad + X_f(X'X)^{-1} X'_f . \end{aligned}$$

Now we need to show that the sum of the latter three terms is 0.

$$\begin{aligned} & X_f(X'X)^{-1}X'_f - \left[X_f(X'X + X'_fX_f)^{-1}X'_f + X_f(X'X + X'_fX_f)^{-1}X'_fX_f(X'X)^{-1}X'_f \right] \\ &= X_f(X'X)^{-1}X'_f - X_f \left[(X'X + X'_fX_f)^{-1} + (X'X + X'_fX_f)^{-1}X'_fX_f(X'X)^{-1} \right] X'_f . \end{aligned}$$

Now it suffices to show that

$$(X'X + X'_fX_f)^{-1} + (X'X + X'_fX_f)^{-1}X'_fX_f(X'X)^{-1} = (X'X)^{-1} .$$

Multiplying both sides by $X'X + X'_fX_f$, we get

$$\begin{aligned} I + X'_fX_f(X'X)^{-1} &= (X'X + X'_fX_f)(X'X)^{-1} \\ &= I + (X'_fX_f)(X'X)^{-1} . \end{aligned}$$

This proves Identity 2.1.

Identity 2.2

$$(I + X_f(X'X)^{-1}X'_f)X_f(X'_fX_f + X'X)^{-1}X'X = X_f .$$

Proof:

Taking $\left[(X'_fX_f + X'X)^{-1}X'X\right]^{-1}$ of both sides, we get

$$\begin{aligned}(I + X_f(X'X)^{-1}X'_f)X_f &= X_f(X'X)^{-1}(X'_fX_f + X'X) \\ &= X_f(X'X)^{-1}X'_fX_f + X_f .\end{aligned}$$

The left hand side is

$$(I + X_f(X'X)^{-1}X'_f)X_f = X_f + X_f(X'X)^{-1}X'_fX_f .$$

Thus Identity 2.2 is established.

Now we have

$$\begin{aligned}
 & \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} + z' (I - X_f (X_f' X_f + X' X)^{-1} X_f') z \\
 = & \hat{\beta}' \Lambda_f' X' X \hat{\beta} + (z - \mu_z)' (I - X_f (X_f' X_f + X' X)^{-1} X_f') (z - \mu_z) \\
 & - \mu_z' (I - X_f (X_f' X_f + X' X)^{-1} X_f') \mu_z ,
 \end{aligned}$$

where

$$\begin{aligned}
 \mu_z &= (I - X_f (X_f' X_f + X' X)^{-1} X_f')^{-1} \left(X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \right) \\
 &= \left(I + X_f (X' X)^{-1} X_f' \right) \left(X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \right) \quad (\text{Identity 1}) \\
 &= X_f \hat{\beta} \quad (\text{Identity 2})
 \end{aligned}$$

Thus $\mu_z = X_f \hat{\beta}$, and we have

$$\begin{aligned}
 & \hat{\beta}' \Lambda_f' X' X \hat{\beta} + (z - X_f \hat{\beta})' \left(I - X_f (X_f' X_f + X' X)^{-1} X_f' \right) (z - X_f \hat{\beta}) \\
 & - \hat{\beta}' X_f' (I - X_f (X_f' X_f + X' X)^{-1} X_f') X_f \hat{\beta} .
 \end{aligned}$$

Now

$$\begin{aligned}
 & \hat{\beta}' \Lambda_f' X' X \hat{\beta} - \hat{\beta}' X_f' \left(I - X_f (X_f' X_f + X' X)^{-1} X_f' \right) X_f \hat{\beta} \\
 &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - \hat{\beta}' X_f' X_f \hat{\beta} + \hat{\beta}' X_f' X_f \Lambda_f \hat{\beta} \\
 &= \hat{\beta}' \left[\Lambda_f' X' X - X_f' X_f + X_f' X_f \Lambda_f \right] \hat{\beta}.
 \end{aligned}$$

Now observe that

$$\begin{aligned}
 & \Lambda_f' X' X - X_f' X_f + X_f' X_f \Lambda_f \\
 &= X_f' X_f (X_f' X_f + X' X)^{-1} X' X - X_f' X_f \\
 &\quad + X_f' X_f (X_f' X_f + X' X)^{-1} X_f' X_f.
 \end{aligned}$$

Claim 2.1

$$X_f' X_f (X_f' X_f + X' X)^{-1} X_f' X_f + X_f' X_f (X_f' X_f + X' X)^{-1} X' X = X_f' X_f .$$

To see this, left multiply both sides by $(X_f' X_f)^{-1}$, which yields

$$(X_f' X_f + X' X)^{-1} X' X + (X_f' X_f + X' X)^{-1} X_f' X_f = I .$$

Now multiply both sides by $X_f' X_f + X' X$, which yields

$$X' X + X_f' X_f = X_f' X_f + X' X .$$

Thus

$$\Lambda_f' X' X - X_f' X_f + X_f' X_f \Lambda_f = 0 .$$

Finally, we have

$$\begin{aligned}
 p(z|X_f, y) &\propto \int_0^\infty \tau^{\frac{n+q-p}{2}-1} \\
 &\times \exp \left\{ -\frac{\tau}{2} \left[(n-p)s^2 + (z - X_f \hat{\beta})' \left(I - X_f (X_f' X_f + X' X)^{-1} X_f' \right) (z - X_f \hat{\beta}) \right] \right\} d\tau \\
 &\propto \left[(n-p)s^2 + (z - X_f \hat{\beta})' \left(I - X_f (X_f' X_f + X' X)^{-1} X_f' \right) (z - X_f \hat{\beta}) \right]^{-\frac{n-p+q}{2}} \\
 &\propto \left[1 + \frac{1}{(n-p)s^2} (z - X_f \hat{\beta})' \left(I + X_f (X' X)^{-1} X_f' \right)^{-1} (z - X_f \hat{\beta}) \right]^{-\frac{n-p+q}{2}}.
 \end{aligned}$$

Thus

$$Z|X_f, y \sim S_q \left(n-p, X_f \hat{\beta}, s^2 (I + X_f (X' X)^{-1} X_f') \right).$$

Theorem 2.7

Consider the linear model in (2.1) and (2.2) and suppose

$$\beta | \tau \sim N_p(\mu_0, \tau^{-1} \Sigma_0)$$

$$\tau \sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right).$$

Let Z be a $q \times 1$ future vector of observations taken at X_f , with

$$Z = X_f \beta + \epsilon_f, \quad \epsilon_f \sim N_q(0, \sigma^2 I).$$

Then

$$Z | X_f, y \sim S_q \left(n + \delta_0, X_f \tilde{\beta}, \tilde{s}^2 (I + X_f (\Sigma_0^{-1} + X' X)^{-1} X_f') \right),$$

where

$$\tilde{s}^2 = (n + \delta_0)^{-1} \left[Y'(I - M)Y + (\hat{\beta} - \mu_0)'(\Lambda' X' X)(\hat{\beta} - \mu_0) + \gamma_0 \right],$$

$$\tilde{\beta} = \Lambda \mu_0 + (I - \Lambda) \hat{\beta},$$

$$\Lambda = (X' X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1}.$$

Proof:

We have

$$\begin{aligned}
 p(z|X_f, y) &\propto \int_0^{\infty} \int_{-\infty}^{\infty} \tau^{\frac{n+p+q+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}[\gamma_0 + (n-p)s^2]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + (\beta - \mu_0)' \Sigma_0^{-1}(\beta - \mu_0)]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[z - X_f \beta]'[z - X_f \beta]\right\} d\beta d\tau \\
 \\
 &\propto \int_0^{\infty} \int_{-\infty}^{\infty} \tau^{\frac{n+p+q+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}[\gamma_0 + (n-p)s^2]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[(\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) + (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0)]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[z - X_f \beta]'[z - X_f \beta]\right\} d\beta d\tau \\
 \\
 &= \int_0^{\infty} \int_{-\infty}^{\infty} \tau^{\frac{n+p+q+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}[(n + \delta_0)\tilde{s}^2]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[(\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[z - X_f \beta]'[z - X_f \beta]\right\} d\beta d\tau \\
 \\
 &= \int_0^{\infty} \int_{-\infty}^{\infty} \tau^{\frac{n+p+q+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}[(n + \delta_0)\tilde{s}^2]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}[(\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]\right\} \\
 &\quad \times \exp\left\{-\frac{\tau}{2}\left[(\beta - \hat{\beta}_z)'(X_f'X_f)(\beta - \hat{\beta}_z) + z'(I - M_{X_f})z\right]\right\} d\beta d\tau .
 \end{aligned}$$

Now

$$\begin{aligned}
 & (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) \\
 & + (\beta - \hat{\beta}_z)'(X_f'X_f)(\beta - \hat{\beta}_z) + z'(I - M_{x_f})z \\
 = & (\tilde{\beta} - \hat{\beta}_z)'(\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z \\
 & + (\beta - \tilde{\beta})' \left(X_f'X_f + X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}),
 \end{aligned}$$

where

$$\tilde{\Lambda} = \left(X_f'X_f + X'X + \Sigma_0^{-1} \right) X_f'X_f.$$

This is clear from the previous proof.

Thus, we have

$$\begin{aligned}
 & \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+p+q+\delta_0}{2}-1} \exp \left\{ -\frac{\tau}{2} [(n + \delta_0)\tilde{s}^2] \right\} \\
 & \times \exp \left\{ -\frac{\tau}{2} [(\beta - \tilde{\beta})'(X_f'X_f + X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})] \right\} \\
 & \times \exp \left\{ -\frac{\tau}{2} [(\tilde{\beta} - \hat{\beta}_z)'(\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z] \right\} d\beta d\tau \\
 = & \int_0^\infty \tau^{\frac{n+q+\delta_0}{2}-1} \exp \left\{ -\frac{\tau}{2} [(n + \delta_0)\tilde{s}^2] \right\} \\
 & \times \exp \left\{ -\frac{\tau}{2} [(\tilde{\beta} - \hat{\beta}_z)'(\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z] \right\} d\tau.
 \end{aligned}$$

Following the earlier proof, we have

$$\begin{aligned}
 & (\tilde{\beta} - \hat{\beta}_z)'(\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z \\
 = & (z - X_f\tilde{\beta})'\left(I - X_f(X_f'X_f + X'X + \Sigma_0^{-1})^{-1}X_f'\right)(z - X_f\tilde{\beta}) \\
 = & (z - X_f\tilde{\beta})'\left(I + X_f(X'X + \Sigma_0^{-1})^{-1}X_f'\right)^{-1}(z - X_f\tilde{\beta}) .
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(z|X_f, y) & \propto \int_0^\infty \tau^{\frac{n+q+\delta_0}{2}-1} \exp\left\{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]\right\} \\
 & \quad \times \exp\left\{-\frac{\tau}{2}\left[(z - X_f\tilde{\beta})'\left(I + X_f(X'X + \Sigma_0^{-1})^{-1}X_f'\right)^{-1}(z - X_f\tilde{\beta})\right]\right\} d\tau \\
 & \propto \left[(n+\delta_0)\tilde{s}^2 + (z - X_f\tilde{\beta})'\left(I + X_f(X'X + \Sigma_0^{-1})^{-1}X_f'\right)^{-1}(z - X_f\tilde{\beta})\right]^{-\frac{m}{2}} \\
 & \propto \left[1 + \frac{1}{(n+\delta_0)\tilde{s}^2}(z - X_f\tilde{\beta})'\left(I + X_f(X'X + \Sigma_0^{-1})^{-1}X_f'\right)^{-1}(z - X_f\tilde{\beta})\right]^{-\frac{m}{2}} ,
 \end{aligned}$$

where $m = n + \delta_0 + q$. Thus

$$z|X_f, y \sim S_q\left(n + \delta_0, X_f\tilde{\beta}, \tilde{s}^2\left(I + X_f(X'X + \Sigma_0^{-1})^{-1}X_f'\right)\right).$$

Note that the noninformative case can be obtained from the informative case by formally setting

$$\delta_0 = -p, \quad \Sigma_0^{-1} = 0, \quad \gamma_0 = 0.$$

Square completion in n dimensions

Suppose $x = (x_1, \dots, x_n)'$ is a $n \times 1$ vector, A is a $n \times n$ nonsingular matrix, and b is an $n \times 1$ vector. Then

$$x'Ax + b'x = \left(x + \frac{A^{-1}b}{2} \right)' A \left(x + \frac{A^{-1}b}{2} \right) - \frac{b'A^{-1}b}{4}.$$

This is the multivariate analog of the one dimensional square completion given earlier. This result is very useful in computing multivariate normal integrals.

Combining quadratic forms

Suppose x is an $n \times 1$ vector, μ_1 and μ_2 are $n \times 1$ vectors, and A_1 and A_2 are $n \times n$ matrices such that $A_1 + A_2$ is nonsingular. Then

$$\begin{aligned} & (x - \mu_1)' A_1 (x - \mu_1) + (x - \mu_2)' A_2 (x - \mu_2) \\ &= (x - \mu^*)' (A_1 + A_2) (x - \mu^*) + \mu_1' A_1 \mu_1 + \mu_2' A_2 \mu_2 - \mu^* (A_1 + A_2) \mu^*, \end{aligned}$$

where

$$\mu^* = (A_1 + A_2)^{-1} (A_1 \mu_1 + A_2 \mu_2).$$

We can generalize this result to combining m quadratic forms, We have

$$\begin{aligned} & (x - \mu_1)' A_1 (x - \mu_1) + \cdots + (x - \mu_m)' A_m (x - \mu_m) \\ &= (x - \mu^*)' B (x - \mu^*) + \sum_{i=1}^m \mu_i' A_i \mu_i - \mu^* B \mu^*, \end{aligned}$$

where $B = \sum_{i=1}^m A_i$ and $\mu^* = B^{-1} (\sum_{i=1}^m A_i \mu_i)$.

Informative Prior Elicitation

Prior elicitation involves the choice of prior parameters once the functional form of the prior is chosen. The functional form of the prior is usually not a difficult task. The functional form of the prior can be facilitated by examining the support of the parameter space as well as interpreting the parameter.

For example, if

$\Theta = \{\theta : -\infty < \theta < \infty\}$, a prior with R^1 as the support is suitable, such as a normal or t . Examples include a normal mean or regression coefficients.

$\Theta = \{\theta : 0 < \theta < \infty\}$, a gamma prior may be suitable. Examples include precision parameters as in a normal model, a Poisson mean, or a scale parameter of an exponential model.

$\Theta = \{\theta : 0 < \theta < 1\}$, a beta prior may be suitable. Examples include probability parameters as in a binomial model.

Thus the functional form can be specified by examining the parameter space and interpreting the parameter. There is also a trade-off with practicality and functional complexity.

For example, a t prior is computationally more difficult to work with than a normal prior. A mixture of normals, i.e.,

$$p_1 \pi_1(\theta) + p_2 \pi_2(\theta),$$

is typically more difficult than a t and so on.

The problem of **prior elicitation** is typically concerned with the specification of the prior hyperparameters once a functional form of the prior has been chosen.

Example 2.1: Normal Priors

The normal prior is useful when $\Theta = R^p$. Suppose we take

$$\theta \sim N_p(\mu_0, \Sigma_0).$$

How do we pick μ_0 and Σ_0 ? μ_0 is a $p \times 1$ vector and Σ_0 is a $p \times p$ matrix. Thus, we have to specify $p + \frac{p(p+1)}{2}$ numbers. The normal prior is a very flexible prior for a parameter space having R^p as its support. Many shapes of unimodal densities can be obtained by varying the choices of μ_0 and Σ_0 . For example, when $p = 1$, $\theta \sim N_1(\mu_0, \sigma_0^2)$.

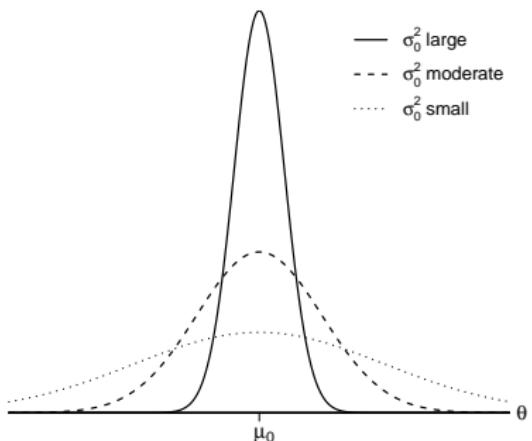


Figure 2.1:

The hyperparameter μ_0 is a location parameter that shifts the prior distribution, and σ_0^2 either flattens or sharpens the distribution. We see that the normal prior is a very flexible prior for parameters whose support is the real line.

Example 2.2: Gamma Prior

The gamma prior is useful when $\Theta = \{\theta : 0 < \theta < \infty\}$. It is usually used in modeling variance (or precision) parameters, such as a normal precision parameter $\theta = 1/\sigma^2$, or a Poisson mean $\theta > 0$. The gamma distribution is very flexible. If we take

$$\theta \sim \text{gamma}(\delta_0, \gamma_0) ,$$

then by varying δ_0, γ_0 one can obtain many different shapes of densities.

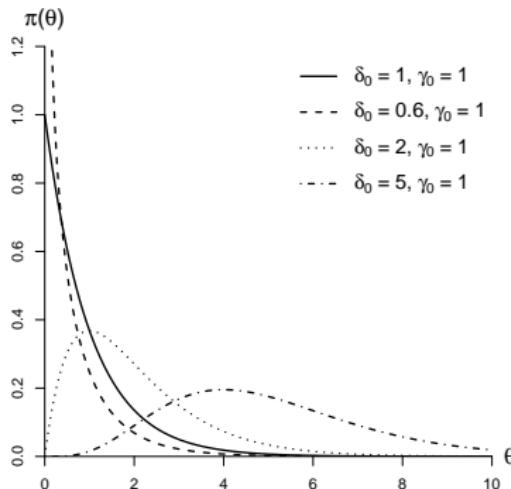
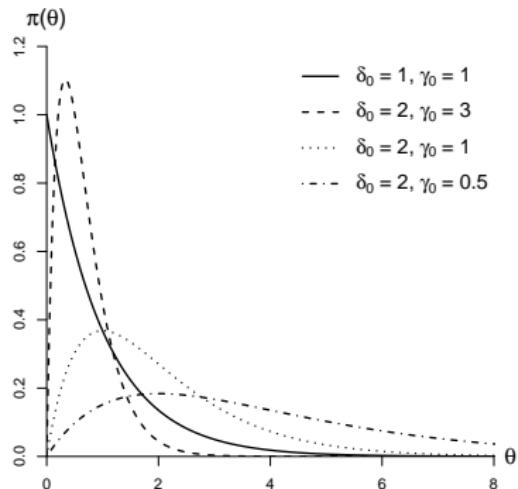


Figure 2.2:

We see that we can generate many shapes as we vary (δ_0, γ_0) .

Example 2.3: Beta Prior

A beta prior is useful when $\Theta = \{\theta : 0 < \theta < 1\}$. For example, θ may be a binomial probability.

For example, if $x | \theta \sim \text{Bin}(N, \theta)$, $0 < \theta < 1$, a beta prior for θ is a reasonable and flexible prior. We write

$$\theta \sim \text{beta}(\alpha, \lambda), \quad \alpha > 0, \quad \lambda > 0.$$

As we vary the hyperparameters, many different shapes of densities can be obtained.

From Figure 2.3, we see the various shapes of densities that can be obtained by varying α and λ .

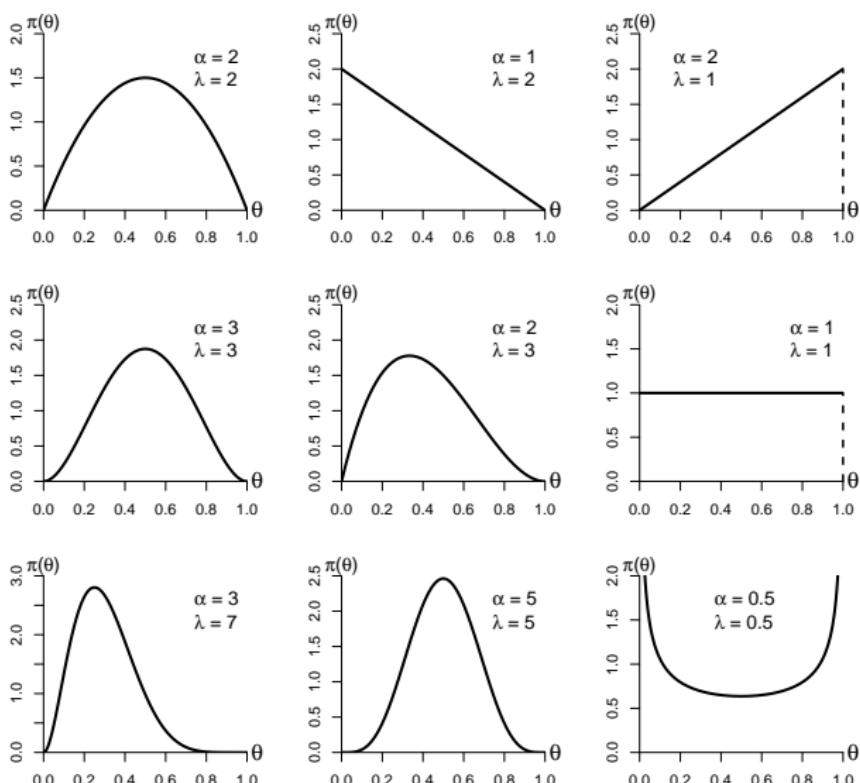


Figure 2.3:

Informative Prior Elicitation for the Linear Model

Now consider the linear model

$$\begin{aligned} Y &= X\beta + \epsilon, \\ \epsilon &\sim N_n(0, \sigma^2 I), \end{aligned}$$

where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, and ϵ is $n \times 1$.

Our parameters are (β, τ) , where $\tau = 1/\sigma^2$. The conjugate informative prior specification for (β, τ) is

$$\begin{aligned} \beta | \tau &\sim N_p(\mu_0, \tau^{-1} \Sigma_0), \\ \tau &\sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

How do we specify (μ_0, Σ_0) and (δ_0, γ_0) in practice?

1. The investigator has previous experience with such data, and thus obtains “estimates” of $(\mu_0, \Sigma_0, \gamma_0, \delta_0)$ from literature or previous similar studies.
2. The investigator has raw data from a previous similar study with which to elicit $(\mu_0, \Sigma_0, \gamma_0, \delta_0)$.

Suppose $D_0 = (n_0, Y_0, X_0)$ is the raw data from a previous study similar to the current study. We will call D_0 the *historical data*.

In this case, one can take

$$\mu_0 = (X_0' X_0)^{-1} X_0' Y_0, \quad X_0 \text{ is } n_0 \times p$$

$$\Sigma_0 = a_0^{-1} (X_0' X_0)^{-1}, \quad \text{where } a_0 \text{ is a specified scalar}$$

$$\frac{\delta_0}{\gamma_0} = \left(\frac{Y_0'(I - M_0)Y_0}{n_0 - p} \right)^{-1}$$

$$\frac{2\delta_0}{\gamma_0^2} = c_0 \left(\frac{Y_0'(I - M_0)Y_0}{n_0 - p} \right)^{-1},$$

where $M_0 = X_0(X_0' X_0)^{-1} X_0'$ and c_0 is a specified scalar. Having previous data $D_0 = (n_0, Y_0, X_0)$ makes the elicitation **semi-automatic**.

Caution: In any Bayesian analysis, **sensitivity analyses** must be conducted by varying the choices of the prior hyperparameters and examining the impact of these choices on the inference. This is a very important and crucial part of any Bayesian analysis using informative priors. It is **never** a good idea to do a Bayesian analysis based on informative priors using just one set of chosen hyperparameters. One must examine sensitivity and robustness of the model to various choices of prior hyperparameters, and these results should be reported.

In addition, it is always a good idea to compare informative prior results with results based on noninformative priors to examine the differences. Here the noninformative prior serves as a benchmark for comparison. For the normal linear model, we can construct a noninformative prior from the conjugate specification by taking

$$\mu_0 = 0, \quad \Sigma_0 = \sigma_0^2 I, \quad \text{and let } \sigma_0^2 \rightarrow \infty,$$

$$\gamma_0 = \delta_0 = 0.$$

Such choices of prior hyperparameters lead to an analysis based on noninformative priors.

A more sophisticated informative prior can be constructed as follows. Suppose we have a previous similar study which yielded data $D_0 = (n_0, Y_0, X_0)$. The likelihood function of (β, τ) based on the previous study is given by

$$L(\beta, \tau | Y_0) = (2\pi)^{-\frac{n_0}{2}} \tau^{n_0/2} \exp \left\{ -\frac{\tau}{2} (Y_0 - X_0\beta)'(Y_0 - X_0\beta) \right\} .$$

For the prior of $\beta | \tau$, we weight $L(\beta, \tau | Y_0)$ by taking it to a power a_0 , where $0 \leq a_0 \leq 1$.

Thus

$$\begin{aligned}\pi(\beta \mid \tau, a_0) &\propto [L(\beta, \tau \mid Y_0)]^{a_0} \\&= (2\pi)^{-\frac{n_0 a_0}{2}} \tau^{\frac{n_0 a_0}{2}} \exp \left\{ -\frac{a_0 \tau}{2} (Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) \right\} \\&\propto \exp \left\{ -\frac{a_0 \tau}{2} (\beta - \hat{\beta}_0)' (X_0' X_0) (\beta - \hat{\beta}_0) \right\}\end{aligned}$$

where $\hat{\beta}_0 = (X_0' X_0)^{-1} X_0' Y_0$.

Thus $\beta \mid \tau, a_0 \sim N_p \left(\hat{\beta}_0, a_0^{-1} \tau^{-1} (X_0' X_0)^{-1} \right)$.

For τ , we specify $\tau \sim \text{gamma} \left(\frac{\delta_0}{2}, \frac{\gamma_0}{2} \right)$. Thus

$$\begin{aligned}\pi(\beta, \tau | a_0) &\propto \exp \left\{ -\frac{a_0 \tau}{2} (\beta - \hat{\beta}_0)' (X_0' X_0) (\beta - \hat{\beta}_0) \right\} \\&\quad \times \tau^{\delta_0/2-1} \exp \left\{ -\frac{\gamma_0 \tau}{2} \right\}.\end{aligned}$$

Remark 2.1

We can also treat a_0 as **random** and take $a_0 \sim \text{beta}(\alpha, \lambda)$, and thus the joint prior would be $\pi(\beta, \tau, a_0) = \pi(\beta, \tau | a_0)\pi(a_0)$. Taking a_0 random gives similar results to a_0 fixed.

Remark 2.2

These priors are very useful in model selection problems since they make the elicitation scheme semi-automatic. They are called *power priors*.

Such priors have been discussed by

Ibrahim and Laud (1994, JASA)

Laud and Ibrahim (1995, JRSS-B)

Laud and Ibrahim (1996, Biometrika)

Ibrahim, Ryan, and Chen (1998, JASA)

Chen, Ibrahim, and Yiannoutsos (1999, JRSS-B)

Ibrahim and Chen (2000, Statistical Science)

Ibrahim et al. (2015, Statistics in Medicine)

Remark 2.3

The full power prior for (β, τ) when τ is unknown is

$$\begin{aligned}\pi(\beta, \tau) &\propto L(\beta, \tau | D_0)^{a_0} \pi_0(\beta, \tau) \\ &\propto \tau^{n_0 a_0 / 2} \exp \left\{ -\frac{a_0 \tau}{2} (Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) \right\} \pi_0(\beta, \tau).\end{aligned}$$

$\pi_0(\beta, \tau)$ is called the *initial prior*.

If we take

$$\begin{aligned}\pi_0(\beta, \tau) &\propto \pi_0(\beta) \pi_0(\tau) \\ &\propto \tau^{\delta_0 / 2 - 1} \exp \{-\gamma_0 \tau / 2\},\end{aligned}$$

then we have

$$\pi(\beta, \tau) = \frac{\tau^{\frac{n_0 a_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} (Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) \right\} \tau^{\delta_0 / 2 - 1} \exp \{ -\gamma_0 \tau / 2 \}}{\int \int \tau^{\frac{n_0 a_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} (Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) \right\} \tau^{\delta_0 / 2 - 1} \exp \{ -\gamma_0 \tau / 2 \} d\beta d\tau}. \quad (2.3)$$

Let $s_0^2 = \frac{Y_0(I - M_0)Y_0}{n_0 - p}$. Then

$$\begin{aligned}
 C &= \int \int \tau^{\frac{n_0 a_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} (Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) \right\} \tau^{\delta_0/2 - 1} \exp \{-\gamma_0 \tau/2\} d\beta d\tau \\
 &= \int \int \tau^{\frac{n_0 a_0 + \delta_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} \left[(n_0 - p)s_0^2 + (\beta - \hat{\beta}_0)' (X_0' X_0) (\beta - \hat{\beta}_0) + \frac{\gamma_0}{a_0} \right] \right\} d\beta d\tau \\
 &= \int (2\pi)^{p/2} (a_0 \tau)^{-p/2} |X_0' X_0|^{-1/2} \tau^{\frac{n_0 a_0 + \delta_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} \left[(n_0 - p)s_0^2 + \gamma_0/a_0 \right] \right\} d\tau \\
 &= (2\pi)^{p/2} |X_0' X_0|^{-1/2} a_0^{-p/2} \int_0^\infty \tau^{\frac{n_0 a_0 + \delta_0 - p}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} \left[(n_0 - p)s_0^2 + \gamma_0/a_0 \right] \right\} d\tau \\
 &= \frac{(2\pi)^{p/2} |X_0' X_0|^{-1/2} a_0^{-p/2} \Gamma \left(\frac{n_0 a_0 + \delta_0 - p}{2} \right)}{\left[\frac{a_0}{2} ((n_0 - p)s_0^2 + \gamma_0/a_0) \right]^{\frac{n_0 a_0 + \delta_0 - p}{2}}}.
 \end{aligned}$$

Thus

$$\pi(\beta, \tau) = \frac{1}{C} \left\{ \tau^{\frac{n_0 a_0 + \delta_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} \left[(Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) + \frac{\gamma_0}{a_0} \right] \right\} \right\}.$$

Remark 2.4

It is more convenient to work with the “conditional” power prior;

$$\begin{aligned}\beta|\tau, a_0 &\sim N_p \left(\hat{\beta}_0, a_0^{-1} \tau^{-1} (X_0' X_0)^{-1} \right) \\ \tau &\sim \text{gamma}(\delta_0/2, \gamma_0/2).\end{aligned}$$

Remark 2.5

The conditional power prior = full power prior when τ is known.

When $\delta_0 = 2$, $\gamma_0 = 0$, we have

$$\pi(\beta, \tau) = \frac{L(\beta, \tau | Y_0)^{a_0}}{\int \int L(\beta, \tau | Y_0)^{a_0} d\beta d\tau}. \quad (2.4)$$

Now

$$\begin{aligned} & \int \int L(\beta, \tau | Y_0)^{a_0} d\beta d\tau \\ C^* &= \frac{(2\pi)^{p/2} |X_0' X_0|^{-1/2} a_0^{-p/2} \Gamma \left(\frac{n_0 a_0 + 2 - p}{2} \right)}{\left[\frac{a_0}{2} (n_0 - p) s_0^2 \right]^{\frac{n_0 a_0 + 2 - p}{2}}} < \infty\end{aligned}$$

so that $\pi(\beta, \tau)$ is proper.

Remark 2.6

$\pi(\beta, \tau)$ in (2.4) is an informative prior in β and an informative prior in τ .
 The marginal prior on τ from (2.4) is

$$\begin{aligned}\pi^*(\tau) &= \int \pi(\beta, \tau) d\beta \\ &= \frac{1}{C^*} \int \tau^{n_0 a_0 / 2} \exp \{-a_0 \tau / 2\} \left[(n_0 - p) s_0^2 + (\beta - \hat{\beta}_0)' (X_0' X_0) (\beta - \hat{\beta}_0) \right] d\beta \\ &= \frac{1}{C^*} \tau^{n_0 a_0 / 2} \exp \left\{ -\frac{a_0 \tau (n_0 - p) s_0^2}{2} \right\} (a_0 \tau)^{-p/2} (2\pi)^{p/2} |X_0' X_0|^{-1/2} \\ &= \frac{\tau^{\frac{n_0 a_0 - p}{2}} \exp \left\{ -\frac{a_0 \tau (n_0 - p) s_0^2}{2} \right\}}{\Gamma \left(\frac{n_0 a_0 + 2 - p}{2} \right)} \left(\frac{a_0 (n_0 - p) s_0^2}{2} \right)^{\frac{n_0 a_0 + 2 - p}{2}} \\ &= \text{gamma} \left(\frac{n_0 a_0 - p + 2}{2}, \frac{a_0 (n_0 - p) s_0^2}{2} \right). \\ E(\tau) &= \frac{\left(\frac{n_0 a_0 - p + 2}{2} \right)}{\left(\frac{a_0 (n_0 - p) s_0^2}{2} \right)} = \frac{n_0 a_0 - p + 2}{a_0 (n_0 - p) s_0^2}.\end{aligned}$$

Under the prior (2.3), that is,

$$\pi(\beta, \tau) = \frac{1}{C} \left\{ \tau^{n_0 a_0 / 2} \exp \left\{ -\frac{a_0 \tau}{2} (Y_0 - X_0 \beta)' (Y_0 - X_0 \beta) \right\} \tau^{\delta_0 / 2 - 1} \exp \left\{ -\gamma_0 \tau / 2 \right\} \right\},$$

the marginal prior of τ is

$$\begin{aligned} \pi^{**}(\tau) &= \int \pi(\beta, \tau) d\beta \\ &= \text{gamma} \left(\frac{n_0 a_0 + \delta_0 - p}{2}, \frac{a_0 (\frac{\gamma_0}{a_0} + (n_0 - p) s_0^2)}{2} \right). \end{aligned}$$

Remark 2.7

Picking $\pi_0(\tau)$ to be a gamma distribution in (2.3) leads to a closed form for the joint prior of (β, τ) in (2.3).

The marginal prior of β from (2.3) is

$$\begin{aligned}
 \pi^{**}(\beta) &= \int \pi(\beta, \tau) d\tau \\
 &= \frac{1}{C} \int \tau^{\frac{n_0 a_0 + \delta_0}{2} - 1} \exp \left\{ -\frac{a_0 \tau}{2} \left[(n_0 - p)s_0^2 + (\beta - \hat{\beta}_0)'(X_0' X_0)(\beta - \hat{\beta}_0) + \frac{\gamma_0}{a_0} \right] \right\} d\tau \\
 &= \frac{1}{C} \Gamma \left(\frac{n_0 a_0 + \delta_0}{2} \right) \left\{ \frac{a_0}{2} \left((n_0 - p)s_0^2 + (\beta - \hat{\beta}_0)'(X_0' X_0)(\beta - \hat{\beta}_0) + \frac{\gamma_0}{a_0} \right) \right\}^{-\frac{(n_0 a_0 + \delta_0)}{2}} \\
 &\propto \left[1 + \frac{(n_0 a_0 + \delta_0 - p)^{-1}}{(n_0 a_0 + \delta_0 - p)^{-1} \left\{ (n_0 - p)s_0^2 + \frac{\gamma_0}{a_0} \right\} (\beta - \hat{\beta}_0)'(X_0' X_0)(\beta - \hat{\beta}_0)} \right]^{-\frac{n_0 a_0 + \delta_0 - p + p}{2}} \\
 &= \left[1 + \frac{(n_0 a_0 + \delta_0 - p)}{(n_0 a_0 + \delta_0 - p) \left\{ (n_0 - p)s_0^2 + \frac{\gamma_0}{a_0} \right\} (\beta - \hat{\beta}_0)'(X_0' X_0)(\beta - \hat{\beta}_0)} \right]^{-\frac{n_0 a_0 + \delta_0 - p + p}{2}}.
 \end{aligned}$$

Thus, marginally, based on (2.3),

$$\beta \sim S_p \left(n_0 a_0 + \delta_0 - p, \hat{\beta}_0, \frac{a_0(n_0 - p)s_0^2 + \gamma_0}{a_0(n_0 a_0 + \delta_0 - p)} (X_0' X_0)^{-1} \right).$$

With $\delta_0 = 2$, $\gamma_0 = 0$, the marginal prior based on (2.4) is

$$\beta \sim S_p \left(n_0 a_0 + 2 - p, \hat{\beta}_0, \frac{(n_0 - p)s_0^2}{n_0 a_0 + 2 - p} (X_0' X_0)^{-1} \right).$$

The conditional prior for $\beta|\tau$ from (2.3) is

$$\begin{aligned}\pi^{**}(\beta|\tau) &\propto \exp\left\{-\frac{a_0\tau}{2}(Y_0 - X_0\beta)'(Y_0 - X_0\beta)\right\} \\ &\propto \exp\left\{-\frac{a_0\tau}{2}\left[(n_0 - p)s_0^2 + (\beta - \hat{\beta}_0)(X_0'X_0)(\beta - \hat{\beta}_0)\right]\right\} \\ &\propto \exp\left\{-\frac{a_0\tau}{2}(\beta - \hat{\beta}_0)(X_0'X_0)(\beta - \hat{\beta}_0)\right\}.\end{aligned}$$

Thus, the induced conditional prior of $\beta|\tau$ from (2.3) is

$$\beta|\tau \sim N_p\left(\hat{\beta}_0, a_0^{-1}\tau^{-1}(X_0'X_0)^{-1}\right).$$

Remark 2.8

A modification of the power prior when a_0 is random is called the **normalized power prior**, and is given by

$$\pi(\theta, a_0) = \pi(\theta|D_0, a_0)\pi(a_0) = \frac{L(\theta|D_0)^{a_0} \pi_0(\theta)}{\int L(\theta|D_0)^{a_0} \pi_0(\theta)d\theta} \pi(a_0), \quad (2.5)$$

where $\pi_0(\theta)$ and $\pi_0(a_0)$ are the initial priors.

We note that (2.5) first specifies a conditional prior distribution for θ given a_0 and then specifies a marginal distribution for a_0 .

For the normalized power prior in (2.5), we must have

$$\int L(\theta|D_0)^{a_0} \pi_0(\theta)d\theta < \infty$$

for $0 < a_0 \leq 1$.

Remark 2.9

Another variation of the power prior is called the **partial borrowing power prior**.

The key idea of the partial borrowing power prior is that the historical data are borrowed only through the common parameters shared in the models for the historical data and the current data.

Thus, strength from the historical data is borrowed through those common parameters and at the same time, the parameters in the power prior are allowed to be different than those in the likelihood function for the current data.

This attractive and flexible feature of the partial borrowing power prior allows the historical data to have different forms (e.g., summary statistics versus individual-level data) or different models than the current data.

Moreover, the partial borrowing power prior can be adapted to the fixed- a_0 , random a_0 , normalized, and commensurate settings.

Remark 2.10

We can extend the power prior to multiple historical datasets. Suppose we have K_0 historical datasets, D_{0k} , $k = 1, \dots, K_0$. Let $D_0 = (D_{01}, \dots, D_{0K_0})$.

We have

$$\pi(\theta) \propto \prod_{k=1}^{K_0} L(\theta | D_{0k})^{a_{0k}} \pi_0(\theta), \quad (2.6)$$

where $\pi_0(\theta)$ is the initial prior for θ , $a_0 = (a_{01}, \dots, a_{0K_0})$, and $0 \leq a_{0k} \leq 1$ for $k = 1, \dots, K_0$.

The prior in (2.6) is attractive since it allows for different a_{0k} 's for different historical datasets, providing a flexible degree of discounting for each historical dataset.

The theoretical and computational properties of (2.6) are similar to those of the single historical dataset case, and (2.6) can also be extended to the variations of the normalized power prior.

Noninformative Prior Elicitation

A prior is said to be noninformative if it satisfies the local uniformity property. Below we discuss several motivations and properties of noninformative priors that may assist in choosing a certain noninformative prior to do the analysis.

Motivation 2.1

If the parameter for which we are specifying the prior is a location parameter to support the real line, then a uniform (improper) prior is reasonable since it will give posterior distributions whose parameters correspond to frequentist point estimates. We saw earlier examples of this for the i.i.d. normal model as well as the linear model. In a model with location and scale parameters, the prior with respect to the scale parameter is typically not a uniform prior. Recall for the normal model,

$$\pi(\beta, \tau) \propto \tau^{-1} .$$

This prior yields a posterior which has parameters that correspond to frequentist point estimates, but is not a uniform prior in τ .

Motivation 2.2

It is desirable to have a noninformative prior that has an attractive interpretation and is easy to compute, such as Jeffreys prior. Another useful way to construct meaningful noninformative priors is to take the corresponding conjugate prior or some other informative prior and choose the parameters in such a way to construct a noninformative prior.

Example 2.4

Suppose $\theta > 0$, and we take

$$\theta \sim \text{gamma}(\delta_0, \gamma_0) .$$

Thus, $\pi(\theta) \propto \theta^{\delta_0 - 1} \exp\{-\gamma_0 \theta\}$. As $\delta_0 \rightarrow 0, \gamma_0 \rightarrow 0$, we get the noninformative prior

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

Example 2.5

Suppose $-\infty < \theta < \infty$, and we take

$$\theta \sim N(\mu_0, \sigma_0^2) .$$

Letting $\sigma_0^2 \rightarrow \infty$, we get a noninformative prior for θ .

Example 2.6

Suppose $0 < \theta < 1$, and we take $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\lambda-1}$. We can construct a noninformative prior by letting $\alpha \rightarrow 0$ and $\lambda \rightarrow 0$, which yields

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

Thus, we can construct noninformative priors by taking a reasonable informative prior for the parameter and choosing the prior hyperparameters in a certain way.

Motivation 2.3

It is desirable to use a noninformative prior that yields a reasonable analytic and computational form for the posterior. Noninformative priors that yield closed forms for the posterior distributions may be attractive.

It is thus desirable to find a noninformative prior that satisfies motivations 2.1, 2.2, and 2.3 simultaneously. For the **normal linear model**,

$$\pi(\beta, \tau) \propto \tau^{-1}$$

satisfies 2.1, 2.2, and 2.3. However other priors may also be used. We can take

$$\pi(\beta, \tau) \propto \tau^{p/2-1} . \quad (\text{Jeffreys prior})$$

This prior yields a closed form posterior and has a nice interpretation. It does not satisfy motivation 2.1 or the latter part of motivation 2.2, i.e., it cannot be obtained from informative priors with certain prior parameters. One can also use

$$\pi(\beta, \tau) \propto 1 .$$

Again, this prior yields a closed form posterior. Also, one can use

$$\beta | \tau \sim N(\mu_0, \tau^{-1} \Sigma_0) ,$$

$$\tau \sim \text{gamma} \left(\frac{\delta_0}{2}, \frac{\gamma_0}{2} \right) ,$$

and take $\mu_0 = 0$, $\Sigma_0 = a_0^{-1} I$, and let $a_0 \rightarrow 0$, $\delta_0 \rightarrow 0$, and $\gamma_0 \rightarrow 0$.

This is one way to construct a proper noninformative prior for the linear model.

Bayes Factors for Linear Models

Theorem 2.8

Consider the linear model

$$Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n)$$

and

$$\begin{aligned}\beta | \tau &\sim N(\mu_0, \tau^{-1} \Sigma_0) \\ \tau &\sim \text{gamma} \left(\frac{\delta_0}{2}, \frac{\gamma_0}{2} \right).\end{aligned}$$

Then, the marginal distribution of Y is given by

$$Y \sim S_n \left(\delta_0, X\mu_0, \frac{\gamma_0}{\delta_0} (I + X\Sigma_0 X') \right),$$

and therefore

$$\begin{aligned} p(y) &= \left(\frac{\Gamma(\frac{n+\delta_0}{2})(\pi\delta_0)^{-n/2} \left(\frac{\gamma_0}{\delta_0}\right)^{-n/2} |I + X\Sigma_0X'|^{-1/2}}{\Gamma(\frac{\delta_0}{2})} \right) \\ &\quad \times \left[1 + \frac{1}{\gamma_0} (Y - X\mu_0)'(I + X\Sigma_0X')^{-1}(Y - X\mu_0) \right]^{-(n+\delta_0)/2}. \end{aligned}$$

Given a model $m \in \mathcal{M}$, we have

$$\begin{aligned} p(y|m) &= \left(\frac{\Gamma\left(\frac{n+\delta_0^{(m)}}{2}\right) (\pi\delta_0^{(m)})^{-n/2} \left(\frac{\gamma_0^{(m)}}{\delta_0^{(m)}}\right)^{-n/2} |I + X^{(m)}\Sigma_0^{(m)}X^{(m)\prime}|^{-1/2}}{\Gamma\left(\frac{\delta_0^{(m)}}{2}\right)} \right) \\ &\quad \times \left[1 + \frac{1}{\gamma_0^{(m)}} (Y - X^{(m)}\mu_0^{(m)})'(I + X^{(m)}\Sigma_0^{(m)}X^{(m)\prime})^{-1}(Y - X^{(m)}\mu_0^{(m)}) \right]^{-\left(n+\delta_0^{(m)}\right)/2}, \end{aligned}$$

where $\mu_0^{(m)}$, $\Sigma_0^{(m)}$, $\delta_0^{(m)}$, and $\gamma_0^{(m)}$, are the hyperparameter values under model m .

The Bayes factor in favor of model m_1 over model m_2 is

$$B_{12} = \frac{p(y|m_1)}{p(y|m_2)}.$$

Derivation of $p(y)$:

Note that

$$p(y) = \int \int p(y|X, \beta, \tau) \pi(\beta|\tau) \pi(\tau) d\beta d\tau.$$

Now using earlier pages of the notes, this integral is proportional to

$$p(y) \propto \left[Y'(I - M)Y + (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0) + \gamma_0 \right]^{-(n+\delta_0)/2},$$

where

$$M = X(X'X)^{-1}X', \quad \Lambda = (X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}.$$

Now observe that

$$\begin{aligned}
 & (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0) \\
 &= \hat{\beta}'(\Lambda'X'X)\hat{\beta} - 2\hat{\beta}'\Lambda'X'X\mu_0 + \mu_0'\Lambda'X'X\mu_0 \\
 &= Y'X(X'X)^{-1}\Lambda'(X'X)(X'X)^{-1}X'Y - 2Y'X(X'X)^{-1}\Lambda'X'X\mu_0 + \mu_0'\Lambda'X'X\mu_0 \\
 &= Y'X(X'X)^{-1}\Lambda'X'Y - 2Y'X(X'X)^{-1}\Lambda'X'X\mu_0 + \mu_0'\Lambda'X'X\mu_0 \\
 &= (Y - X\mu_0)'(X(X'X)^{-1}\Lambda'X')(Y - X\mu_0).
 \end{aligned}$$

Now we need to add

$$\begin{aligned}
 & Y'(I - M)Y + (Y - X\mu_0)'(X(X'X)^{-1}\Lambda'X')(Y - X\mu_0) \\
 &= (Y - \tilde{Y})' [I - M + X(X'X)^{-1}\Lambda'X'] (Y - \tilde{Y}) \\
 &\quad - \tilde{Y}' [I - M + X(X'X)^{-1}\Lambda'X'] \tilde{Y} + \mu_0'\Lambda'X'X\mu_0,
 \end{aligned}$$

where

$$\tilde{Y} = (I - M + X(X'X)^{-1}\Lambda'X')^{-1}(X(X'X)^{-1}\Lambda'X')(X\mu_0).$$

Claim 2.2

$$(I - M + X(X'X)^{-1}\Lambda'X')^{-1} = I + X\Sigma_0X'.$$

Proof:

$$\begin{aligned}
 & [I - M + X(X'X)^{-1}\Lambda'X'][I + X\Sigma_0X'] \\
 &= I - M + (I - M)X\Sigma_0X' + X(X'X)^{-1}\Lambda'X' + X(X'X)^{-1}\Lambda'(X'X)\Sigma_0X' \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + \Lambda'X'X\Sigma_0X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + \Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}X'X\Sigma_0X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + \Sigma_0^{-1}(I - \Lambda)\Sigma_0X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + (I - \Sigma_0^{-1}\Lambda\Sigma_0)X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + X' - \Sigma_0^{-1}\Lambda\Sigma_0X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + X' - \Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}\Sigma_0X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + X' - \Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}X'] \\
 &= I - M + X(X'X)^{-1} [\Lambda'X' + X' - \Lambda'X'] \\
 &= I - M + X(X'X)^{-1}X' \\
 &= I - M + M \\
 &= I.
 \end{aligned}$$

Thus

$$\begin{aligned}
 \tilde{Y} &= (I + X\Sigma_0X')(X(X'X)^{-1}\Lambda'X')X\mu_0 \\
 &= (X(X'X)^{-1}\Lambda'X'X + X\Sigma_0(X'X)(X'X)^{-1}\Lambda'X'X)\mu_0 \\
 &= X((X'X)^{-1}\Lambda'X'X + \Sigma_0\Lambda'X'X)\mu_0 \\
 &= X((X'X)^{-1}\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}X'X + (X'X + \Sigma_0^{-1})^{-1}X'X)\mu_0 \\
 &= X(X'X)^{-1}(\Sigma_0^{-1} + X'X)(X'X + \Sigma_0^{-1})^{-1}X'X\mu_0 \\
 &= X(X'X)^{-1}X'\mu_0 = X\mu_0.
 \end{aligned}$$

Claim 2.3

$$-\tilde{Y}'(I - M + X(X'X)^{-1}\Lambda'X')\tilde{Y} + \mu_0'\Lambda'X'X\mu_0 = 0.$$

Proof:

$$\begin{aligned}
 &-\tilde{Y}'(I - M + X(X'X)^{-1}\Lambda'X')\tilde{Y} \\
 &= -\mu_0'X'(I - M + X(X'X)^{-1}\Lambda'X')X\mu_0 \\
 &= -\mu_0'X'(I - M)X\mu_0 - \mu_0'X'(X(X'X)^{-1}\Lambda'X')X\mu_0 \\
 &= -\mu_0'\Lambda'X'X\mu_0.
 \end{aligned}$$

Thus

$$\begin{aligned} p(y) &\propto \left[\gamma_0 + (Y - X\mu_0)'(I + X\Sigma_0 X')^{-1}(Y - X\mu_0) \right]^{-(n+\delta_0)/2} \\ &\propto \left[1 + \frac{1}{\gamma_0} (Y - X\mu_0)'(I + X\Sigma_0 X')^{-1}(Y - X\mu_0) \right]^{-(n+\delta_0)/2}. \end{aligned}$$

Thus

$$Y \sim S_n \left(\delta_0, X\mu_0, \frac{\gamma_0}{\delta_0} (I + X\Sigma_0 X') \right).$$

Example 2.7: Carcinoma Data

We discuss a practical application which highlights calculations of posterior probabilities and Bayes factors.

Table 2.1: Ages and survival times of 20 carcinoma patients

Age (years)	38	54	37	47	51	48	42	50	45	33
survival time (weeks)	25	45	238	194	16	23	30	16	22	123
Age (years)	46	34	66	44	64	49	56	43	45	40
survival time (weeks)	51	412	45	162	14	72	5	35	30	91

Let

$$\begin{aligned}y^* &= \text{survival time} \\y &= \log(y^*) \\x &= \text{age.}\end{aligned}$$

We consider the linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, 20$$

where

$$\begin{aligned}y_i &= \log(\text{survival time}) \text{ for the } i^{th} \text{ patient,} \\x_i &= \text{age of } i^{th} \text{ patient,}\end{aligned}$$

and $\epsilon_i \sim N(0, \sigma^2)$, $\tau = \frac{1}{\sigma^2}$.

$\log(\text{survival time})$

3.218876 3.806662 5.472271 5.267858 2.772589 3.135494 3.401197 2.772589 3.091042 4.812184
3.931826 6.021023 3.806662 5.087596 2.639057 4.276666 1.609438 3.555348 3.401197 4.510860

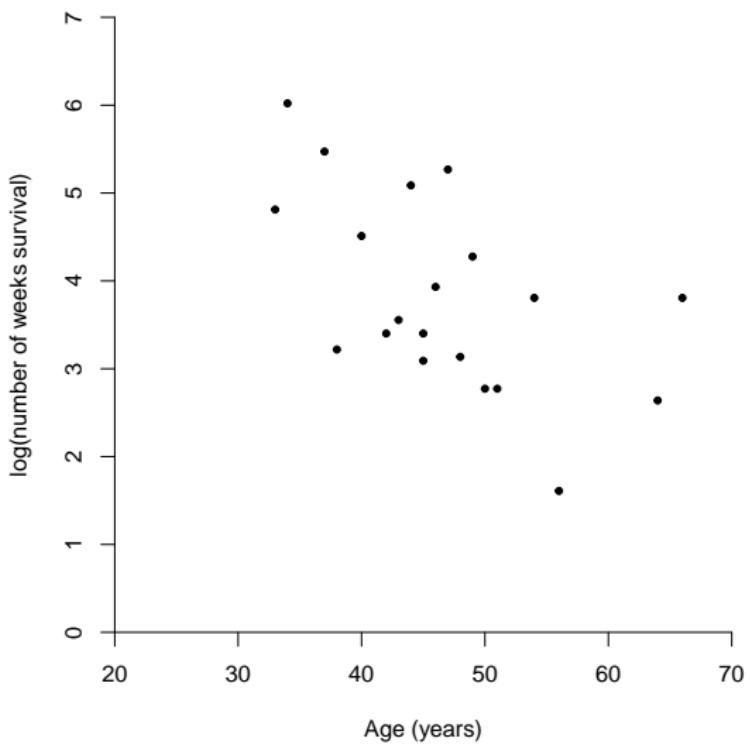


Figure 2.4:

- (a) Consider the case for which $\tau = 1$ and β is unknown. Suppose $\pi(\beta) \propto 1$. Then

$$\beta | y \sim N_2(\hat{\beta}, (X'X)^{-1}), \quad (2.7)$$

where

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y = \begin{pmatrix} 7.16 \\ -.0714 \end{pmatrix}, \\ (X'X)^{-1} &= \begin{pmatrix} 1.54 & -.0391 \\ -.0391 & .000684 \end{pmatrix}.\end{aligned}$$

Compute $P(\beta_1 > 0 | y)$:

From (2.7), we have $\beta_1 | y \sim N(-.0714, .000684)$. Thus

$$\begin{aligned}P(\beta_1 > 0 | y) &= 1 - \Phi\left(\frac{0 + .0714}{\sqrt{.000684}}\right) \\ &= 1 - \Phi(2.73) \\ &= 1 - .9968 = .0032.\end{aligned}$$

(b) Now suppose that τ is unknown and

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Then

$$\beta|y \sim S_2(18, \hat{\beta}, s^2(X'X)^{-1}),$$

$$\text{where } s^2 = \frac{Y'(I-M)Y}{18} = (.9361)^2 = .8762.$$

Compute $P(\beta_1 > 0|y)$:

We note that

$$\beta|y \sim S_2\left(18, \begin{pmatrix} 7.16 \\ -0.0714 \end{pmatrix}, .8762 \begin{pmatrix} 1.54 & -0.0391 \\ -0.0391 & 0.000684 \end{pmatrix}\right),$$

where

$$.8762 \begin{pmatrix} 1.54 & -0.0391 \\ -0.0391 & 0.000684 \end{pmatrix} = \begin{pmatrix} 1.35 & -0.034 \\ -0.034 & 0.00060 \end{pmatrix}.$$

Theorem 2.9

If $x \sim S_p(v, \mu, \Sigma)$, where $x = (x_1, \dots, x_p)$, then the marginal distributions of x are also multivariate t .

Thus if

$$x = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix},$$

where $x^{(1)}$ is $r \times 1$, $x^{(2)}$ is $(p - r) \times 1$,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then $x^{(1)} \sim S_r(v, \mu_1, \Sigma_{11})$.

For our problem, we have

$$\beta_1|y \sim S_1(18, - .0714, (.8762)(.000684)).$$

Thus

$$\begin{aligned} P(\beta_1 > 0|y) &= P\left(t_{18} > \frac{0 + .0714}{\sqrt{(.000684)(.8762)}}\right) \\ &= P(t_{18} > 2.92) \\ &= 1 - P(t_{18} \leq 2.92) \\ &= 1 - .9954 = .0046. \end{aligned}$$

R code for Example 2.7 (a) and (b):

```
x <- c(38,54,37,47,51,48,42,50,45,33,46,34,66,44,64,49,56,43,45,40)
ystar <- c(25,45,238,194,16,23,30,16,22,123,51,412,45,162,14,72,5,35,30,91)
y <- log(ystar)
> y
[1] 3.218876 3.806662 5.472271 5.267858 2.772589 3.135494 3.401197 2.772589 3.091042 4.812184
[11] 3.931826 6.021023 3.806662 5.087596 2.639057 4.276666 1.609438 3.555348 3.401197 4.510860

reg <- lm(y ~ x)
> summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.54916 -0.74633  0.03171  0.70322  1.46689 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.15557   1.16043   6.166 8.03e-06 *** 
x           -0.07137   0.02449  -2.914  0.00926 **  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9361 on 18 degrees of freedom
Multiple R-squared:  0.3205,    Adjusted R-squared:  0.2828 
F-statistic: 8.492 on 1 and 18 DF,  p-value: 0.00926
```

```

int <- rep(1,20)
> int
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

xmat <- cbind(int,x)
> xmat
      int  x
[1,] 1 38
[2,] 1 54
[3,] 1 37
[4,] 1 47
[5,] 1 51
[6,] 1 48
[7,] 1 42
[8,] 1 50
[9,] 1 45
[10,] 1 33
[11,] 1 46
[12,] 1 34
[13,] 1 66
[14,] 1 44
[15,] 1 64
[16,] 1 49
[17,] 1 56
[18,] 1 43
[19,] 1 45
[20,] 1 40

```

```
xxinv <- solve(t(xmat) %*% xmat)
> xxinv
      int           x
int  1.53655531 -0.0319003286
x    -0.03190033  0.0006845564

# Example 2.7 (a)
> 1 - pnorm(0, -0.0714, sqrt(.000684))
[1] 0.003166263

> anova(reg)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
x          1  7.4418  7.4418  8.4916 0.00926 **
Residuals 18 15.7746  0.8764
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

# Example 2.7 (b)
> 1 - pt(2.92, 18)
[1] 0.004570785
```

(c) Suppose we wanted to test

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0,$$

where

$$\begin{aligned} \text{under } H_0 \quad \beta_0 | \tau &\sim N(7, \tau^{-1}) \\ \tau &\sim \text{gamma}(\delta_0/2, \gamma_0/2), \quad \delta_0 = \gamma_0 = .1 \end{aligned}$$

$$\begin{aligned} \text{under } H_1 \quad \beta | \tau &\sim N_2 \left(\begin{pmatrix} 7 \\ -.01 \end{pmatrix}, \tau^{-1} I_2 \right) \\ \tau &\sim \text{gamma}(\delta_0/2, \gamma_0/2), \quad \delta_0 = \gamma_0 = .1 \end{aligned}$$

Let us compute the Bayes factor for this hypothesis. We use the formula derived earlier.

Let

- $m_1 = \text{model under } H_0, \text{ with parameters } (\beta_0), \tau$
- $m_2 = \text{model under } H_1, \text{ with parameters } (\beta_0, \beta_1), \tau.$

We have $\delta_0^{(m)} = \gamma_0^{(m)} = .1$ for both m_1 and m_2 , $X^{(m_1)} = J_{n \times 1}$ where J is the $n \times 1$ vector of ones,

$$X^{(m_2)} = [J \quad x]_{n \times 2} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

$$\mu_0^{(m_1)} = 7, \quad \mu_0^{(m_2)} = \begin{pmatrix} 7 \\ -.01 \end{pmatrix}, \quad \Sigma_0^{(m_1)} = 1, \quad \Sigma_0^{(m_2)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We calculate $p(y|m_1)$ and $p(y|m_2)$ using Theorem 2.8 on pages 221–222.

R code for Example 2.7 (c):

```
n <- 20
x <- c(38,54,37,47,51,48,42,50,45,33,46,34,66,44,64,49,56,43,45,40)
ystar <- c(25,45,238,194,16,23,30,16,22,123,51,412,45,162,14,72,5,35,30,91)
y <- log(ystar)

delta01 <- .1
delta02 <- .1
gamma01 <- .1
gamma02 <- .1

mu01 <- 7
mu02 <- c(7,-.01)

sigma01 <- 1
a1 <- c(1,0)
a2 <- c(0,1)
sigma02 <- cbind(a1,a2)

xm1 <- rep(1,n)
xm1 <- cbind(xm1)
int <- rep(1,n)
xm2 <- cbind(int,x)

term1 <- gamma((n+delta01) / 2) * (pi * delta01)^(-n/2) * (gamma01 / delta01)^(-n/2) /
  gamma(delta01/2)

term2 <- gamma((n+delta02) / 2) * (pi * delta02)^(-n/2) * (gamma02 / delta02)^(-n/2) /
  gamma(delta02/2)
```

```
mat1 <- diag(int, 20, 20) + xm1 %*% sigma01 %*% t(xm1)
mat2 <- diag(int, 20, 20) + xm2 %*% sigma02 %*% t(xm2)

eig1 <- eigen(mat1)$values
eig2 <- eigen(mat2)$values

det1 <- prod(eig1)^(-1/2)
det2 <- prod(eig2)^(-1/2)

quad1 <- t((y - xm1 %*% mu01)) %*% solve(mat1) %*% (y - xm1 %*% mu01)
quad2 <- t((y - xm2 %*% mu02)) %*% solve(mat2) %*% (y - xm2 %*% mu02)

ker1 <- (1 + quad1/gamma01)^(-(n + delta01) / 2)
ker2 <- (1 + quad2/gamma02)^(-(n + delta02) / 2)

pm1 <- term1 * det1 * ker1
pm2 <- term2 * det2 * ker2

b12 <- pm1 / pm2

> b12
[1,] [,1]
[1,] 0.03963679
```

The Bayes factor in favor of model m_1 over model m_2 is $B_{12} = 0.0396$.

The Bayes factor in favor of model m_2 over model m_1 is $B_{21} = 1/0.03963679 = 25.23$.

Theorem 2.10

Consider $Y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, $\tau = 1/\sigma^2$, and τ is known.

Suppose further that

$$\beta \sim N_p(\mu_0, \tau^{-1}\Sigma_0).$$

Then

$$Y \sim N_n(X\mu_0, \tau^{-1}(I + X\Sigma_0 X')),$$

so that

$$\begin{aligned} p(y) &= (2\pi)^{-n/2} \tau^{n/2} |I + X\Sigma_0 X'|^{-1/2} \\ &\quad \times \exp \left\{ -\frac{\tau}{2} (y - X\mu_0)' (I + X\Sigma_0 X')^{-1} (y - X\mu_0) \right\}. \end{aligned}$$

Thus for a given model $m \in \mathcal{M}$,

$$\begin{aligned} p(y|m) &= (2\pi)^{-n/2} \tau_m^{n/2} |I + X^{(m)} \Sigma_0^{(m)} X^{(m)'}|^{-1/2} \\ &\quad \times \exp \left\{ -\frac{\tau_m}{2} (y - X^{(m)} \mu_0^{(m)})' (I + X^{(m)} \Sigma_0^{(m)} X^{(m)'})^{-1} (y - X^{(m)} \mu_0^{(m)}) \right\}, \end{aligned}$$

so that for two models m_1 and m_2 , $B_{12} = \frac{p(y|m_1)}{p(y|m_2)}$.

Bayesian Model Selection

We will first discuss the problem of variable subset selection for the linear model. This is one of the most important problems in statistical inference.

Let \mathcal{M} denote the model space, and suppose the full model has p covariates, so that

$$\begin{aligned} Y &= X\beta + \epsilon, \\ \epsilon &\sim N_n(0, \sigma^2 I), \end{aligned} \tag{2.8}$$

$X_{n \times p}$, $\beta_{p \times 1}$, $\epsilon_{n \times 1}$, $Y_{n \times 1}$. An intercept is always included, so that there are $p + 1$ regression coefficients in all.

Let $m \in \mathcal{M}$. Under model m , we can write (2.8) as

$$Y = X^{(m)}\beta^{(m)} + \epsilon$$

where $X^{(m)}$ is an $n \times k_m$ matrix of covariates of rank k_m , and $\beta^{(m)}$ is a $k_m \times 1$ vector of regression coefficients. Under model m , there are $k_m - 1$ covariates.

To do Bayesian variable selection, we must use **proper priors** on **all** of the model parameters. As discussed earlier, one possible choice of informative priors for the regression coefficients is to take, under model m ,

$$\beta^{(m)} | \tau, a_0 \sim N_{k_m} \left((X_0^{(m)'} X_0^{(m)})^{-1} X_0^{(m)'} Y_0, a_0^{-1} \tau^{-1} (X_0^{(m)}, X_0^{(m)})^{-1} \right),$$

Where $D_0^{(m)} = (n_0, X_0^{(m)}, Y_0)$ is the historical data under model m . Here $X_0^{(m)}$ is an $n_0 \times k_m$ matrix of covariates from the historical data, n_0 is the sample size of the historical data, and Y_0 is the $n_0 \times 1$ response vector from the historical data.

For τ , we take

$$\tau \sim \text{gamma} \left(\frac{\delta_0}{2}, \frac{\gamma_0}{2} \right).$$

We take a_0 **fixed** in order to get closed forms for the posterior and predictive distributions. If a_0 is taken random, we will not get closed forms for the posterior or predictive distributions. We will discuss the case of a_0 random in more detail later, when we develop Gibbs sampling techniques. For now, let us consider a_0 fixed.

Thus, our hyperparameters are

$$(a_0, \delta_0, \gamma_0) .$$

We see that this informative prior elicitation scheme eases the elicitation process a lot since we must specify only 3 numbers for **all** possible models in \mathcal{M} . Thus, the elicitation scheme is semi-automatic in the sense that once $(a_0, \delta_0, \gamma_0)$ are specified, the prior distributions of $\beta^{(m)} | \tau, a_0$ and τ are immediately determined for all possible models.

The parameter a_0 is a precision parameter that weights the historical data. Thus a_0^{-1} is a variance parameter. It is reasonable to restrict $0 \leq a_0 \leq 1$, since it does not make much sense to weight the historical data more than the current data. Values of a_0 close to 1 give high weight to the historical data, whereas values of a_0 close to 0 (i.e., $a_0^{-1} \rightarrow \infty$) give small weight. We see from the construction of the prior that as $a_0^{-1} \rightarrow \infty$, the prior for $\beta^{(m)} | \tau, a_0$ becomes more and more flat.

The parameters (δ_0, γ_0) can also be estimated from the historical data. Since

$$\pi(\tau) \propto \tau^{\delta_0/2 - 1} \exp\left\{-\frac{\gamma_0 \tau}{2}\right\},$$

we have

$$E(\tau) = \frac{\delta_0}{\gamma_0},$$

$$\text{Var}(\tau) = \frac{2\delta_0}{\gamma_0^2}.$$

One possible strategy is to elicit the prior mean and variance of τ , and then solve for (δ_0, γ_0) . For example, we may take

$$\begin{aligned} \frac{\delta_0}{\gamma_0} &= (\text{MSE}_0)^{-1} \\ &= \left[\frac{Y'_0(I - M_0)Y_0}{n_0 - (p + 1)} \right]^{-1}, \end{aligned}$$

where $M_0 = X_0(X'_0 X_0)^{-1} X'_0$ is the orthogonal projection matrix based on the full model with all p covariates.

Then we can choose the prior variance to be a multiple of the prior mean, i.e.,

$$\text{Var}(\tau) = c_0 E(\tau) ,$$

and therefore

$$\frac{2\delta_0}{\gamma_0^2} = c_0 \left[\frac{Y_0'(I - M_0)Y_0}{n_0 - (p + 1)} \right]^{-1} .$$

Small values of c_0 imply a sharp prior about the mean and large values of c_0 imply a flatter prior about the mean. Once c_0 is specified, we solve the two equations for δ_0 and γ_0 . There are other ways of specifying δ_0 and γ_0 . For example, we can take

$$\frac{\delta_0}{\gamma_0} = (\text{MSE}_0)^{-1}$$

and

$$P(\tau < (\text{MSE}_0)^{-1}) = \alpha ,$$

where α is a prespecified probability.

Example 2.8: Carcinoma Data (cont.)

R code to calculate the Bayes factor for specified values of $(a_0, \delta_0, \gamma_0)$:

```
n <- 20
x <- c(38,54,37,47,51,48,42,50,45,33,46,34,66,44,64,49,56,43,45,40)
ystar <- c(25,45,238,194,16,23,30,16,22,123,51,412,45,162,14,72,5,35,30,91)
y <- log(ystar)

n0 <- 15
a0 <- 1.0

y0 <- c(3.806304,4.093267,4.038489,4.534852,5.052120,3.573818,4.537654,2.295017,
       4.130641,4.128961,4.066597,3.277398,3.373287,4.497034,4.029427)

x01 <- c(47,43,44,37,30,50,37,68,42,42,43,54,53,37,44)

int0 <- rep(1,n0)

x0mat1 <- cbind(int0)
x0mat2 <- cbind(int0,x01)

delta01 <- .1
delta02 <- .1
gamma01 <- .1
gamma02 <- .1

mu01 <- lm(y0 ~ int0 - 1)$coefficients
mu02 <- lm(y0 ~ x01)$coefficients
sigma01 <- a0^(-1) * solve(t(x0mat1) %*% x0mat1)
sigma02 <- a0^(-1) * solve(t(x0mat2) %*% x0mat2)

xm1 <- cbind( rep(1,n) )
int <- rep(1,n)
xm2 <- cbind(int,x)
```

```

term1 <- gamma((n + delta01) / 2) * (pi * delta01)^(-n/2) * (gamma01 / delta01)^(-n/2)
term2 <- gamma((n + delta02) / 2) * (pi * delta02)^(-n/2) * (gamma02 / delta02)^(-n/2)

mat1 <- diag(int, 20, 20) + xm1 %*% sigma01 %*% t(xm1)
mat2 <- diag(int, 20, 20) + xm2 %*% sigma02 %*% t(xm2)

eig1 <- eigen(mat1)$values
eig2 <- eigen(mat2)$values

det1 <- prod(eig1)^(-1/2)
det2 <- prod(eig2)^(-1/2)

quad1 <- t((y - xm1 %*% mu01)) %*% solve(mat1) %*% (y - xm1 %*% mu01)
quad2 <- t((y - xm2 %*% mu02)) %*% solve(mat2) %*% (y - xm2 %*% mu02)

ker1 <- (1 + quad1/gamma01)^(-(n + delta01) / 2)
ker2 <- (1 + quad2/gamma02)^(-(n + delta02) / 2)

pm1 <- term1 * det1 * ker1
pm2 <- term2 * det2 * ker2

b12 <- pm1 / pm2

> b12
[1,] [,1]
[1,] 0.02984364

```

With $a_0 = 1.0$, the Bayes factor in favor of model m_1 over model m_2 is $B_{12} = 0.0298$.

a_0	Bayes factor (BF)	1/BF
.1	.0771	12.97
.3	.0469	21.32
.5	.0381	26.24
.7	.0336	29.76
1.0	.0298	33.56

```
> mu01
    int0
3.962324

> mu02
(Intercept)      x01
7.19584334 -0.07228433

> sigma01
    int0
int0 0.06666667

> sigma02
    int0      x01
int0  1.81138107 -0.0390025575
x01  -0.03900256  0.0008718903
```

R function to calculate the Bayes factor when specifying a_0 and solving for (δ_0, γ_0) :

```
bf.fun <- function(a0, c0, c1){

  # Current data
  n <- 20
  x <- c(38,54,37,47,51,48,42,50,45,33,46,34,66,44,64,49,56,43,45,40)
  ystar <- c(25,45,238,194,16,23,30,16,22,123,51,412,45,162,14,72,5,35,30,91)
  y <- log(ystar)
  n0 <- 15
  xm0 <- rep(1,n)
  xm1 <- cbind(xm0,x)

  # Historical data (y0 is already on log scale)
  y0 <- c(3.806304,4.093267,4.038489,4.534852,5.052120,3.573818,4.537654,2.295017,
         + 4.130641,4.128961,4.066597,3.277398,3.373287,4.497034,4.029427)
  x01 <- c(47,43,44,37,30,50,37,68,42,42,43,54,53,37,44)

  # Create intercept for historical data
  int0 <- rep(1,n0)

  # Create X matrices for historical data
  ximat <- cbind(int0,x01)
  x0mat <- int0

  # First, use historical data to get prior for beta
  mu0 <- solve(t(x0mat) %*% x0mat) %*% t(x0mat) %*% y0
  sigma0 <- (1/a0) * solve(t(x0mat) %*% x0mat)
  mu1 <- solve(t(ximat) %*% ximat) %*% t(ximat) %*% y0
  sigma1 <- (1/a0) * solve(t(ximat) %*% ximat)
```

```
## Get prior on tau from historical data

# Calculate mse0
m0 <- x0mat %*% solve(t(x0mat) %*% x0mat) %*% t(x0mat)
proj0 <- diag(n0) - m0
mse0 <- t(y0) %*% proj0 %*% y0 / (length(y0) - 1)
invmse0 <- solve(mse0)
m1 <- x1mat %*% solve(t(x1mat) %*% x1mat) %*% t(x1mat)
proj1 <- diag(n0) - m1
mse1 <- t(y0) %*% proj1 %*% y0 / (length(y0) - 2)
invmse1 <- solve(mse1)

# Solve two equations with two unknowns for delta0 and gamma0
# Equation 1: delta0 / gamma0 = invmse0
# Equation 2: 2 * delta0 / gamma0^2 = c0 * invmse0
gamma0 <- 2/c0
delta0 <- invmse0 * gamma0
gamma1 <- 2/c1
delta1 <- invmse1 * gamma1

term0 <- gamma((n + delta0) / 2) * (pi * delta0)^(-n/2) * (gamma0/delta0)^(-n/2) / gamma(delta0 / 2)
term1 <- gamma((n + delta1) / 2) * (pi * delta1)^(-n/2) * (gamma1/delta1)^(-n/2) / gamma(delta1 / 2)

mat0 <- diag(n) + xm0 %*% sigma0 %*% t(xm0)
mat1 <- diag(n) + xm1 %*% sigma1 %*% t(xm1)

eig0 <- eigen(mat0)$values
eig1 <- eigen(mat1)$values

det0 <- prod(eig0)^(-1/2)
det1 <- prod(eig1)^(-1/2)
```

```
quad0 <- t((y - xm0 %*% mu0)) %*% solve(mat0) %*% (y - xm0 %*% mu0)
quad1 <- t((y - xm1 %*% mu1)) %*% solve(mat1) %*% (y - xm1 %*% mu1)

ker0 <- (1 + quad0/gamma0)^(-(n + delta0) / 2)
ker1 <- (1 + quad1/gamma1)^(-(n + delta1) / 2)

pm0 <- term0 * det0 * ker0
pm1 <- term1 * det1 * ker1

bf01 <- pm0 / pm1

bf01

}
```

Bayesian variable selection results in computing **posterior model probabilities** for all models $m \in \mathcal{M}$.

$$p(m \mid \text{data}) = \frac{p(\text{data} \mid m) p(m)}{\sum_{m \in \mathcal{M}} p(\text{data} \mid m) p(m)}$$

or

$$p(m \mid y) = \frac{p(y \mid m) p(m)}{\sum_{m \in \mathcal{M}} p(y \mid m) p(m)} .$$

For the normal-gamma prior discussed above,

$$\begin{aligned} p(y \mid m) &= \int \int p(y \mid X^{(m)}, \beta^{(m)}, \tau) \pi(\beta^{(m)} \mid \tau, a_0) \pi(\tau) d\beta^{(m)} d\tau \\ &\propto \int \int \tau^{\frac{n+\delta_0-1}{2}} \exp \left\{ -\frac{\tau}{2} (Y - X^{(m)} \beta^{(m)})' (Y - X^{(m)} \beta^{(m)}) \right\} \\ &\quad \times \tau^{k_m/2} \exp \left\{ -\frac{a_0 \tau}{2} (\beta^{(m)} - \mu_0^{(m)})' (X_0^{(m)'} X_0^{(m)}) (\beta^{(m)} - \mu_0^{(m)}) \right\} \\ &\quad \times \exp \left\{ -\frac{\gamma_0 \tau}{2} \right\} d\beta^{(m)} d\tau , \end{aligned}$$

where $\mu_0^{(m)} = (X_0^{(m)'} X_0^{(m)})^{-1} X_0^{(m)'} Y_0$.

Using previous results, it can be shown that

$$p(y | m) \propto [Y'(I - M^{(m)})Y + (\hat{\beta}^{(m)} - \mu_0^{(m)})'(\Lambda^{(m)'}X^{(m)'}X^{(m)})(\hat{\beta}^{(m)} - \mu_0^{(m)}) + \gamma_0]^{-\frac{\nu}{2}}$$

where $\nu = n + \delta_0$,

$$M^{(m)} = X^{(m)}(X^{(m)'}X^{(m)})^{-1}X^{(m)'}$$

$$\Lambda^{(m)} = [X^{(m)'}X^{(m)} + a_0 X_0^{(m)'}X_0^{(m)}]^{-1}(a_0 X_0^{(m)'}X_0^{(m)}) .$$

This derivation is based on the formulas derived earlier. Now $p(y | m)$ is a density in Y , so we must fix up the formula above. It turns out that $p(y | m)$ is a multivariate t distribution. I will leave it as an exercise for you to complete the derivation above.

Prior on the Model Space

Now that we are able to compute $p(y | m)$ for all $m \in \mathcal{M}$, we need to specify a prior on the model space \mathcal{M} . That is, we need to specify $p(m)$ for all $m \in \mathcal{M}$, such that $\sum_{m \in \mathcal{M}} p(m) = 1$.

A common choice used by many is a uniform prior, that is

$$p(m) = 2^{-p}.$$

This may not be desirable if p is large. If one has historical data $D_0^{(m)} = (n_0, X_0^{(m)}, Y_0)$, we can use the historical data to construct $p(m)$. A possible choice is to take

$$p(m) = \frac{\int \int p(y_0 | X_0^{(m)}, \beta^{(m)}, \tau) \pi_0(\beta^{(m)}, \tau) d\beta^{(m)} d\tau}{\sum_{m \in M} \int \int p(y_0 | X_0^{(m)}, \beta^{(m)}, \tau) \pi_0(\beta^{(m)}, \tau) d\beta^{(m)} d\tau},$$

where $\pi_0(\beta^{(m)}, \tau)$ is the **initial prior** distribution for $\beta^{(m)}, \tau$, that is, the prior for $(\beta^{(m)}, \tau)$ **before** $D_0^{(m)} = (n_0, X_0^{(m)}, Y_0)$ is observed.

We can take noninformative proper priors here, i.e.,

$$\pi_0(\beta^{(m)}, \tau) = \pi_0(\beta^{(m)} | \tau) \pi_0(\tau),$$

$$\pi_0(\beta^{(m)} | \tau) \propto \exp \left\{ -\frac{\tau d_0}{2} \beta^{(m)'} \beta^{(m)} \right\},$$

$$\pi_0(\tau) \propto \tau^{\frac{\delta_0^*}{2} - 1} \exp \left\{ -\frac{\gamma_0^*}{2} \tau \right\},$$

and choose

$$d_0 \rightarrow 0, \quad \delta_0^* \rightarrow 0, \quad \gamma_0^* \rightarrow 0.$$

Thus

$$p(m) \equiv p(m | y_0)$$

with $\pi_0(\beta^{(m)}, \tau)$ denoting the **initial prior** distribution for $(\beta^{(m)}, \tau)$.

With $\pi_0(\beta^{(m)}, \tau)$ as given above, $p(m)$ will have a closed form for the linear model. This provides us with an informative prior elicitation for the model space \mathcal{M} .

Remark 2.11

Recall from earlier developments, we took the conditional power prior to be

$$\pi(\beta|\tau, a_0) \propto [L(\beta, \tau|Y_0)]^{a_0}.$$

We can generalize this to

$$\pi(\beta|\tau, a_0) \propto [L(\beta, \tau|Y_0)]^{a_0} \pi_0(\beta)$$

where $\pi_0(\beta)$ is the initial prior for β .

If $a_0 = 1$, then $\pi(\beta|\tau, a_0)$ is a coherent Bayesian update of $\pi_0(\beta)$. That is, $\pi(\beta|\tau, a_0)$ is the posterior distribution of β based on historical data.

Informative Prior Elicitation

For the historical carcinoma data in Example 2.8, we have

$$\begin{aligned}\mu^{(m)} &= (X_0^{(m)'} X_0^{(m)})^{-1} X_0^{(m)'} y_0, \\ \Sigma_0^{(m)} &= a_0^{-1} (X_0^{(m)'} X_0^{(m)})^{-1}, \\ D_0^{(m)} &= (n_0, X_0^{(m)}, y_0), \\ \mu_0^{(m_1)} &= 3.962, \\ \mu_0^{(m_2)} &= \begin{pmatrix} 7.196 \\ -0.072 \end{pmatrix}, \\ \Sigma_0^{(m_1)} &= a_0^{-1} 0.067, \\ \Sigma_0^{(m_2)} &= a_0^{-1} \begin{pmatrix} 1.811 & -0.039 \\ -0.039 & 0.00087 \end{pmatrix}.\end{aligned}$$

Elicitation from Estimates

Suppose we have single model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

and we have estimates summarized from a previous study, given by

$$\text{Var}(x_0) = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (x_{0i} - \bar{x}_0)^2,$$

$$\text{Var}(y_0) = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (y_{0i} - \bar{y}_0)^2,$$

$$\text{Cov}(x_0, y_0) = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (x_{0i} - \bar{x}_0)(y_{0i} - \bar{y}_0).$$

We can construct priors from these estimates as follows: Let $\mu_0 = \begin{pmatrix} \mu_{01} \\ \mu_{02} \end{pmatrix}$

denote the prior mean vector of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

We can take

$$\mu_{02} = \frac{\text{Cov}(x_0, y_0)}{\text{Var}(x_0)}, \quad (2.9)$$

$$\mu_{01} = \bar{y}_0 - \mu_{02} \bar{x}_0, \quad (2.10)$$

$$\Sigma_0 = a_0^{-1} (X_0' X_0)^{-1} = a_0^{-1} \begin{pmatrix} n_0 & n_0 \bar{x}_0 \\ n_0 \bar{x}_0 & \sum_{i=1}^{n_0} x_{0i}^2 \end{pmatrix}^{-1}, \quad (2.11)$$

$$X_0 = \begin{pmatrix} 1 & x_{011} \\ \vdots & \vdots \\ 1 & x_{0n1} \end{pmatrix}_{n_0 \times 2}.$$

Note that $\begin{pmatrix} \mu_{01} \\ \mu_{02} \end{pmatrix}$ are constructed as least squares estimates, and Σ_0 is a multiple of the Fisher information matrix constructed from the estimates.

Example 2.9: Carcinoma Data (cont.)

Suppose we have the following summaries as historical data.

$$\begin{aligned} n_0 &= 15, \\ \bar{x}_0 &= 44.73, \\ \bar{y}_0 &= 3.96, \\ \text{Var}(x_0) &= 81.92, \\ \text{Var}(y_0) &= .428, \\ \text{Cov}(x_0, y_0) &= -5.922. \end{aligned}$$

Then, using (2.9),(2.10),(2.11), we can obtain the prior mean and covariance matrix of β as $\beta \sim N_2(\mu_0, \Sigma_0)$.

$$\begin{aligned}\mu_0 &= \begin{pmatrix} \mu_{01} \\ \mu_{02} \end{pmatrix}, \\ \mu_{02} &= -5.922/81.924 = -.0722, \\ \mu_{01} &= 3.96 - (-.0722)(44.73) \\ &= 7.19.\end{aligned}$$

Thus, the prior mean of β is

$$\begin{aligned}\mu_0 &= \begin{pmatrix} 7.19 \\ -.0722 \end{pmatrix}, \\ \Sigma_0 &= a_0^{-1} \begin{pmatrix} 15 & 670.95 \\ 670.95 & 31158.5 \end{pmatrix}^{-1} = a_0^{-1} \begin{pmatrix} 1.811 & -.039 \\ -.039 & .00087 \end{pmatrix}, \\ n_0 \bar{x}_0 &= (15)(44.73) = 670.95, \\ \sum_{i=1}^{n_0} x_{0i}^2 &= (n_0 - 1)\text{Var}(x_0) + n_0 \bar{x}_0^2 \\ &= (14)(81.92) + 15(44.73)^2 = 31158.5.\end{aligned}$$

Criterion Based Methods

As mentioned earlier, two advantages of Bayesian criterion based methods for model selection are that we do not need to specify a prior over the model space \mathcal{M} , and we **can** use **improper** priors for the regression parameters. For variable subset selection in the linear model, the general form of many variable selection criteria for two nested models $m \subset m_0$ is given by

$$\Lambda(a) = \lambda - a(k_{m_0} - k_m) ,$$

where λ denotes the likelihood ratio statistic for testing m against m_0 , i.e.,

$$\lambda = -2 \log \left[\frac{\max p(y | \beta^{(m)}, \tau)}{\max p(y | \beta^{(m_0)}, \tau)} \right] ,$$

where $k_{m_0} = \text{rank}(X^{(m_0)})$, and $k_m = \text{rank}(X^{(m)})$.

The term a quantifies a penalty for overfitting. If $a \geq 1$, then smaller models are favored over more complex models.

Some Choices for a

<u>a</u>	<u>Criterion</u>
2	AIC (equivalent to Mallow's C_p)
$\log(n)$	BIC
3/2	Local Bayes Factor (Smith and Spiegelhalter, 1980, JRSS-B)

Other values of a have been proposed by Ibrahim and Laud (1994, JASA), Laud and Ibrahim (1995, JRSS-B), as well as others also mentioned in the article by Laud and Ibrahim (1995, JRSS-B).

Bayesian variable selection is only a special case of the general problem of Bayesian model selection. For the general model selection problem, we do not need nested models. For example, we can entertain models that involve Box-Cox power transformations on Y , such as

$$\log(Y) = X\beta + \varepsilon$$

$$\sqrt{Y} = X\beta + \varepsilon$$

$$Y^{(\alpha)} = X\beta + \varepsilon, \quad \alpha \text{ is chosen.}$$

We can also consider variable subset selection. Thus, if we consider the above transformations along with the variable subset selection for each, our model space consists of $r2^p$ models, where r is the number of transformations on Y . Then posterior probability computations proceed as before.

The L measure

Ibrahim and Laud (1994, JASA), Laud and Ibrahim (1995, JRSS-B), Gelfand and Ghosh (1998, Biometrika) and Ibrahim, Chen, and Sinha (2001, Statistica Sinica) discuss a Bayesian criterion called the *L measure* for model selection.

Consider an experiment that yields the data $y = (y_1, \dots, y_n)$. Denote the joint sampling density of the y_i 's by $p(y|\theta)$, where θ is a vector of indexing parameters.

We allow the y_i 's to be fully observed, right censored, or interval censored. In the right censored case, y_i may be a failure time or a censored time. In the interval censored case, we only observe the interval $[a_{l_i}, a_{r_i}]$ in which y_i occurred.

Let $z = (z_1, \dots, z_n)$ denote future values of an imagined replicate experiment. That is, z is a future response vector with the same sampling density as $y|\theta$. Thus, y and z have a common design matrix X .

The imagined replicate experiment makes y and z directly comparable, and independent a priori. It seems clear that good models should make predictions close to what has been observed for an identical experiment.

With this notion in mind, Ibrahim and Laud (1994) defined their statistic as the expected squared Euclidean distance between y and z ,

$$L = E[(z - y)'(z - y)] , \quad (2.12)$$

where the expectation is taken with respect to the predictive distribution of $z|y$,

$$p(z|y) = \int p(z|\theta) p(\theta|y) d\theta .$$

Here θ denotes the vector of indexing parameters, $p(z|\theta)$ is the sampling distribution of the future vector z , and $p(\theta|y)$ denotes the posterior distribution of θ .

Straightforward algebra shows that L can be written as

$$L = \sum_{i=1}^n (\text{Var}(z_i|y) + (E(z_i|y) - y_i)^2) .$$

Thus L can be written as a sum of two terms, one involving the predictive variances and the other term is like a bias term involving the squared difference between the predictive means and the observed data.

For variable selection, under model m , we write

$$L_m = \sum_{i=1}^n (\text{Var}(z_i|y) + (E(z_i|y) - y_i)^2) , \quad (2.13)$$

where

$$p(z|X^{(m)}, y) = \int p(z|X^{(m)}, \theta^{(m)}) p(\theta^{(m)}|X^{(m)}, y) d\theta^{(m)} .$$

For the linear model $\theta^{(m)} = (\beta^{(m)}, \tau)$. Assuming

$$\pi(\beta^{(m)}, \tau) \propto \tau^{-1},$$

the resulting predictive distribution is

$$(z|X^{(m)}, y) \sim S_n(n - k_m, M_m y, s_m^2(I + M_m)) ,$$

where

$$s_m^2 = (n - k_m)^{-1} y'(I - M_m)y ,$$

and $M_m = X^{(m)}(X^{(m)'} X^{(m)})^{-1} X^{(m)'}$.

In this case,

$$L_m = 2(n - 1)(n - k_m - 2)^{-1} q_m, \quad (2.14)$$

where $q_m = y'(I - M_m)y$ is the residual sum of squares under model m .

Now, we consider a special case of the power priors, with $X_0^{(m)} = X^{(m)}$ and $n_0 = n$. In this case, y_0 is viewed as a prior prediction for y rather than historical data. The L measure can also be obtained for a general $X_0^{(m)}$ and n_0 , and thus can be computed when we have historical data.

We take $\beta^{(m)} | \tau$ to be normally distributed

$$\beta^{(m)} | \tau \sim N_{k_m}(\mu^{(m)}, \tau^{-1} a_0^{-1} \Sigma_m),$$

where $\mu^{(m)} = (X^{(m)'} X^{(m)})^{-1} X^{(m)'} y_0$ and $\Sigma_m = (X^{(m)'} X^{(m)})^{-1}$.

The prior distribution for τ is taken to be a gamma distribution with hyperparameters $(\delta_0/2, \gamma_0/2)$.

Under this power prior, we have

$$(z | X^{(m)}, y) \sim S_n(n + \delta_0, \eta_m, s_m^2(I + (1 - b_0)M_m)),$$

where

$$\begin{aligned} b_0 &= a_0 / (1 + a_0), \\ \eta_m &= M_m(b_0 y_0 + (1 - b_0)y), \\ s_m^2 &= (n + \delta_0)^{-1}(q_m + \gamma p_m + \gamma_0), \\ q_m &= y'(I - M_m)y, \\ p_m &= (y - y_0)'M_m(y - y_0). \end{aligned}$$

The L measure under model m is now given by

$$L_m = (1 + \lambda_m)q_m + b_0(b_0 + \lambda_m)p_m + \lambda_m\gamma_0, \quad (2.15)$$

where $\lambda_m = \frac{n+(1-b_0)k_m}{n+\delta_0-2}$.

We see that L_m from Equation (2.15) is a linear function of q_m and p_m . The quantity q_m is the squared length of the projection of the data onto the error space of model m , i.e., the error sum of squares for model m . The quantity p_m represents a penalty for a bad prior guess at y .

The L measure under noninformative priors can be obtained by formally setting $b_0 = 0$, $\delta_0 = -k_m$, and $\gamma_0 = 0$.

If

$$\beta | \tau \sim N_p(\mu_0, \tau^{-1} a_0^{-1} \Sigma_0) ,$$

where $\mu_0 = (X'_0 X_0)^{-1} X'_0 y_0$, $\Sigma_0 = (X'_0 X_0)^{-1}$, and $\tau \sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)$, then

$$z | y \sim S_n(n + \delta_0, X(\Lambda\mu_0 + (I_p - \Lambda)\hat{\beta}), \tilde{s}^2(I_n + X(a_0 \Sigma_0^{-1} + X'X)^{-1} X')) , \quad (2.16)$$

where $\Lambda = a_0(X'X + a_0\Sigma_0^{-1})^{-1}\Sigma_0^{-1}$, I_p denotes the $p \times p$ identity matrix, $\hat{\beta} = (X'X)^{-1}X'Y$, and

$$\tilde{s}^2 = (n + \delta_0)^{-1} \left[Y'(I - M)Y + (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0) + \gamma_0 \right] .$$

Exercise 2.1

Derive the L measure based on (2.16).

Smith and Spiegelhalter (1980) present a general form for variable selection criteria. For two nested models $m \subset m_0$, this form is given by

$$\Lambda(a) = \lambda - a (k_{m_0} - k_m) ,$$

where λ denotes the likelihood ratio statistic and a quantifies a penalty for overfitting. They further point out that if $a \geq 1$, then smaller models are favored over more complex models. Clayton, Geisser, and Jennings (1986) also mention that this is a sensible property, especially for prediction problems.

Under noninformative priors, it can be shown that

$$2n \log \left(\frac{L_m}{L_{m_0}} \right) = \lambda - a_L (k_{m_0} - k_m) ,$$

where

$$a_L = \frac{n}{k_{m_0} - k_m} \log \left(\frac{n - k_m - 2}{n - k_{m_0} - 2} \right) .$$

Thus $a_L > 1$ for all n , and it decreases to 1 as $n \rightarrow \infty$.

A widely accepted non-Bayesian criterion for variable selection in the linear model is Mallows's C_p , which is equivalent to Akaike's AIC.

A standard Bayesian criterion is Schwarz's BIC. Criteria based on a predictive Bayesian distribution include those of Geisser and Eddy (1979) and San Martini and Spezzaferri (1984, 1986).

The deviance information criterion (DIC), proposed by Spiegelhalter et al. (2002), is given by

$$\text{DIC} = D(\bar{\theta}) + 2p_D,$$

where

$$p_D = \bar{D}_\theta - D(\bar{\theta}),$$

$$\bar{\theta} = E(\theta|D),$$

$$\bar{D}_\theta = E\{D(\theta)|D\},$$

$$D(\theta) = \text{deviance} = -2 \{\log(L(\theta)) - \log(L(y))\}.$$

DIC is becoming popular and easy to implement.

Model Checking for the Linear Model

When examining the adequacy of the linear model, we typically examine residuals, the MSE, R^2 , diagonal elements of $M = X(X'X)^{-1}X'$, and so on. There are Bayesian counterparts to these frequentist model checking techniques.

Statistics that are computed to check the adequacy of a model are called **diagnostics**.

For example, the ordinary residuals are defined as

$$\hat{\epsilon}_i = y_i - x_i' \hat{\beta} .$$

The studentized residuals are defined as

$$t_i = \frac{y_i - x_i' \hat{\beta}}{\sqrt{\text{var}(y_i - x_i' \hat{\beta})}} = \frac{y_i - x_i' \hat{\beta}}{\hat{\sigma} \sqrt{1 - m_{ii}}} .$$

To check the adequacy of a linear model, one often checks for outliers or influential points. An outlier is a “strange” value in the response space, and an influential point is a point that affects the regression fit a lot. An influential point may be a strange point in the predictor space, response space, or both.

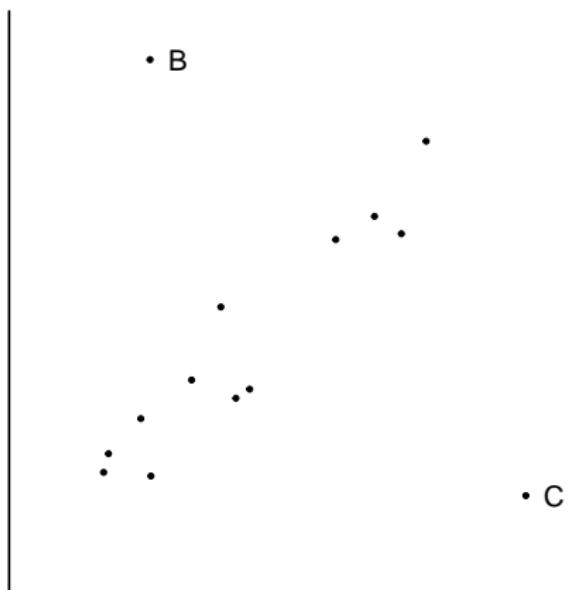


Figure 2.5: C is an influential point and B is an outlier.

Detection of outliers and influential points typically relies on examining residuals.

There have been many Bayesian methods proposed for outlier detection and residual analysis. See Geisser (1993) for several methods and several references. Here, we discuss a method by Chaloner and Brant (1988, *Biometrika*). Consider the usual linear model

$$Y = X\beta + \epsilon ,$$

$$\epsilon \sim N_n(0, \sigma^2 I) .$$

To detect which observations are outliers, define the probability p_i to be

$$p_i = P(|\epsilon_i| > k\sigma | y) .$$

That is, p_i is the **posterior probability** that the i^{th} observation is an outlier. Here k is a constant to be discussed shortly.

Let $\Phi(\cdot)$ denote the standard normal c.d.f. Further, let

$$\begin{aligned} z_{1i} &= \frac{k - \hat{\epsilon}_i \sqrt{\tau}}{\sqrt{m_{ii}}} \\ z_{2i} &= \frac{-k - \hat{\epsilon}_i \sqrt{\tau}}{\sqrt{m_{ii}}} , \end{aligned}$$

where $\tau = 1/\sigma^2$, and m_{ii} is the i^{th} diagonal element of $M = X(X'X)^{-1}X'$.

Then, we have

$$p_i = \int \{1 - \Phi(z_{1i}) + \Phi(z_{2i})\} p(\tau \mid y) \ d\tau .$$

We note that p_i is computed assuming the joint noninformative prior $\pi(\beta, \tau) \propto \tau^{-1}$.

To decide on the magnitude of p_i , we compare it to the prior probability

$$P(|\epsilon_i| > k\sigma) = 2\Phi(-k) .$$

Choice of k

The value of k can be chosen so that the prior probability of no outliers in the entire data set is large, say .95. This yields the formula

$$k = \Phi^{-1} \left(.5 + .5(.95)^{\frac{1}{n}} \right) .$$

Thus any observation with posterior probability of being an outlier larger than the prior probability of $2\Phi(-k)$ would be suspect.

Note that the prior probability of no outliers is

$$\begin{aligned} & \prod_{i=1}^n P(|\epsilon_i| \leq k\sigma) \\ &= [\Phi(k) - \Phi(-k)]^n \\ &= [\Phi(k) - (1 - \Phi(k))]^n \\ &= [2\Phi(k) - 1]^n . \end{aligned}$$

Setting $(2\Phi(k) - 1)^n = .95$, we get

$$2\Phi(k) - 1 = (.95)^{\frac{1}{n}},$$

$$\begin{aligned}\Phi(k) &= \frac{1 + (.95)^{\frac{1}{n}}}{2} \\ &= .5 + .5(.95)^{\frac{1}{n}},\end{aligned}$$

$$k = \Phi^{-1} \left(.5 + .5(.95)^{\frac{1}{n}} \right).$$

We get the following table for various values of n .

n	k
20	3.0
50	3.34
100	3.5
1000	4.0

Example 2.10: Carcinoma Data (cont.)

R code to calculate the Chaloner and Brant statistic:

```
n <- 20
pp <- 2
x <- c(38,54,37,47,51,48,42,50,45,33,46,34,66,44,64,49,56,43,45,40)
ystar <- c(25,45,238,194,16,23,30,16,22,123,51,412,45,162,14,72,5,35,30,91)
y <- log(ystar)
k <- rep(3,n)
lmf <- lm(y ~ x)           # linear models object
lms <- summary(lmf)
lmi <- lm.influence(lmf)   # influence object
epshat <- residuals(lmf)  # residuals
mii <- lmi$hat             # diagonal elements of M
s2 <- (lms$sigma)^2         # MSE
alpha <- (n - pp) / 2      # posterior shape hyperparameter for tau
lambda <- (n - pp) * s2 / 2 # posterior scale hyperparameter for tau
```

```

# Chaloner and Brant statistic function
cb.func <- function(tau, alpha.cb, lambda.cb, k.cb, epshat.cb, m.cb){
  Z1 <- (k.cb - epshat.cb * sqrt(tau)) / sqrt(m.cb)
  Z2 <- (-k.cb - epshat.cb * sqrt(tau)) / sqrt(m.cb)
  p.tau <- lambda.cb^alpha.cb * gamma(alpha.cb)^(-1) * tau^(alpha.cb - 1) * exp(-lambda.cb * tau)
  constant <- 1 - pnorm(Z1) + pnorm(Z2)
  return( constant * p.tau )
}

p <- numeric(n)      # vector to store C&B stats answers
for(i in 1:n){
  p[i] <- integrate(cb.func, lower = 0, upper = Inf,
                     alpha.cb = alpha, lambda.cb = lambda, k.cb = k[i],
                     epshat.cb = epshat[i], m.cb = mi[i])$value
}

# Chaloner and Brant statistic values
> p
[1] 6.374720e-06 9.569446e-16 8.811808e-08 2.366325e-05 9.164626e-15 5.010376e-21 2.651970e-14
[8] 4.864160e-14 4.263526e-14 9.663601e-13 1.029740e-42 2.084487e-04 4.957959e-03 1.067439e-09
[15] 4.049064e-09 8.080064e-20 9.210044e-04 9.172657e-21 1.658440e-22 5.641964e-30

# P(|epsilon_i| > k * sigma) for k = 3.0
> 2 * pnorm(-3.0)
[1] 0.002699796

```

There have been several other methods proposed. These are predictive approaches to outlier detection as discussed by Geisser (1993). Also Pettitt and Smith (1985, *Bayesian Statistics 2*) discuss methods of Bayesian outlier detection.

Let $y_{(-i)}$ denote the response vector with the i^{th} observation deleted. That is, $y_{(-i)} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$. Similarly, let $X_{(-i)}$ denote the $(n - 1) \times p$ covariate matrix **without** the i^{th} row. Thus, the data given by $(Y_{(-i)}, X_{(-i)})$ have the i^{th} case deleted.

Geisser proposes the **Conditional Predictive Ordinate** (CPO) statistic for outlier detection.

The CPO statistic, denoted CPO_i , is the predictive density of the i^{th} observation given $(y_{(-i)}, X_{(-i)})$, evaluated at y_i . That is

$$\text{CPO}_i = p(z_i \mid y_{(-i)}, X_{(-i)})|_{z_i=y_i} .$$

The values of CPO_i give a ranking of the observations, with the most **discordant** having the **smallest value** of CPO_i .

Geisser also defines a p-value for a predictive discordancy test as the tail area of

$$p(z_i \mid y_{(-i)}, X_{(-i)}) .$$

Thus, we compute

$$P[p(z_i \mid y_{(-i)}, X_{(-i)}) \leq \text{CPO}_i] .$$

See Geisser (1993) and the references therein for more details.

Example 2.11: Carcinoma Data (cont.)

R code to calculate the CPO statistic and B statistic:

```
n <- 20
pp <- 2
lmdf <- n - pp          # df of linear model
cpodf <- n - 1 - pp     # df omitting one observation
x <- c(38,54,37,47,51,48,42,50,45,33,46,34,66,44,64,49,56,43,45,40)
ystar <- c(25,45,238,194,16,23,30,16,22,123,51,412,45,162,14,72,5,35,30,91)
y <- log(ystar)
xm <- cbind(rep(1,n), x)

lmf <- lm(y ~ x)          # linear models object
lms <- summary(lmf)
lmi <- lm.influence(lmf)
betahat <- lmf$coefficients
epshat <- residuals(lmf)  # residuals
mii <- lmi$hat            # diagonal elements of M
s2 <- (lms$sigma)^2        # MSE

xpx <- t(xm) %*% xm
xpxinv <- solve(xpx)
xpy <- t(xm) %*% y

cpo <- numeric(n)          # initializing vectors
betai <- numeric(2)
j <- rep(1,n)
```

```

for(i in 1:n){
  xipxiinv <- xpxinv + (xpxinv %*% xm[i,] %*% t(xm[i,]) %*% xpxinv) /
    as.vector( (1 - t(xm[i,]) %*% xpxinv %*% xm[i,]) )
  betai <- (xipxiinv) %*% (xpy - xm[i,] * y[i])
  s2i <- ( (n - pp) * s2 - epshat[i]^2 / (1 - mii[i]) ) / (n - 1 - pp)
  meanzi <- t(xm[i,]) %*% betai
  dispzi <- s2i * (1 + t(xm[i,]) %*% xipxiinv %*% xm[i,])
  tcriti <- (y[i] - meanzi) / sqrt(dispzi)
  cpo[i] <- (1 / sqrt(dispzi)) * dt(tcriti, cpdf)
}

logcpo <- log(cpo)
Bstat <- j %*% logcpo      # calculating B statistic for model adequacy

# CPO statistic
> cpo
[1] 0.14977112 0.33482821 0.21601759 0.10650051 0.28580770 0.32450081 0.28198974
[8] 0.26874209 0.26019347 0.37028220 0.39692267 0.11936725 0.06982171 0.20056752
[15] 0.35104082 0.31833892 0.07814599 0.33634017 0.33563835 0.38145899

# B statistic
> Bstat
[,1]
[1,] -29.26615

```

Computing CPO for the Linear Model

Consider the linear model

$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon &\sim N_n(0, \sigma^2 I) \end{aligned}$$

where $\tau = 1/\sigma^2$, $X_{n \times p}$ of rank p , β is $p \times 1$,

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Let $X_{(i)}$ denote the covariate matrix with the i th row deleted, and let $Y_{(i)}$ denote the response vector with the i th component deleted. That is, $X_{(i)} = X_{(-i)}$ and $Y_{(i)} = Y_{(-i)}$. Let z be an $n \times 1$ vector of future observations taken at X . Then,

$$(z|X_{(i)}, Y_{(i)}) \sim S_n(n - p - 1, X_{(i)}\hat{\beta}_{(i)}, s_{(i)}^2(I + X_{(i)}X'_{(i)})^{-1}X').$$

Therefore,

$$(z_i | X_{(i)}, Y_{(i)}) \sim S_1(n - p - 1, x_i' \hat{\beta}_{(i)}, s_{(i)}^2 (1 + x_i' (X_{(i)}' X_{(i)})^{-1} x_i)),$$

where $\hat{\beta}_{(i)}$ = MLE of β with the i th case deleted, $s_{(i)}^2 = \frac{Y_{(i)}' (I - M_{(i)}) Y_{(i)}}{n - p - 1}$, and $M_{(i)} = X_{(i)} (X_{(i)}' X_{(i)})^{-1} X_{(i)}'$.

Now

$$\text{CPO}_i = p(z | x_{(i)}, y_{(i)})|_{z_i=y_i}.$$

A statistic to asses model adequacy is $B = \sum_{i=1}^n \log(\text{CPO}_i)$.

Recursion Formulas

$$\begin{aligned}(X'_{(i)} X_{(i)})^{-1} &= (X' X - x_i x'_i)^{-1} \\&= (X' X)^{-1} + \frac{(X' X)^{-1} x_i x'_i (X' X)^{-1}}{1 - x'_i (X' X)^{-1} x_i}, \\ \hat{\beta}_{(i)} &= (X'_{(i)} X_{(i)})^{-1} X'_{(i)} Y_{(i)} \\&= (X'_{(i)} X_{(i)})^{-1} [X' Y - x_i Y_i], \\ s^2_{(i)} &= \frac{(n-p)s^2 - \hat{\epsilon}_i^2 / (1 - m_{ii})}{n-p-1}.\end{aligned}$$

Chapter 3:

Bayesian Computational Methods

Bayesian Computational Methods

In most Bayesian inference problems, the posterior distribution of the parameters of interest **does not** have an analytic closed form. That is, the normalizing constant of the posterior distribution does not have a closed form. Recall that the normalizing constant is given by

$$p(x) = \int_{\Theta} p(x | \theta) \pi(\theta) d\theta . \quad (3.1)$$

There are several methods for approximating the posterior distribution, when $p(x)$ cannot be evaluated analytically. We note that the approximation of the posterior distribution boils down to approximating the usually high dimensional integral in (3.1). One of the simplest approximations is a large sample approximation to the posterior distribution obtained by expanding the posterior distribution in a Taylor's series. This approximation is often called the **Bayesian Central Limit Theorem**.

Theorem 3.1: Bayesian Central Limit Theorem

Suppose x_1, \dots, x_n are independent observations from $p(x | \theta)$. Suppose that $\pi(\theta)$ is the prior for θ , which may be improper. Further, suppose that the posterior distribution is proper and its mode exists.

Then as $n \rightarrow \infty$,

$$\theta | x \rightarrow N_p \left[\hat{\theta}_m, H^{-1}(\hat{\theta}_m) \right] ,$$

where $\hat{\theta}_m$ is the posterior mode of θ , obtained by solving

$$\frac{\partial}{\partial \theta_j} \log p^*(\theta | x) = 0 ,$$

where $p^*(\theta | x) = p(x | \theta) \pi(\theta)$ is the kernel of the posterior density (the unnormalized posterior).

$$H(\theta) = - \left[\frac{\partial^2 \log p^*(\theta | x)}{\partial \theta_i \partial \theta_j} \right]$$

is a $p \times p$ matrix. The quantity $-H(\theta)$ is called the **Hessian matrix**. Thus, the asymptotic covariance matrix of θ is the negative of the inverse of the Hessian matrix, evaluated at $\hat{\theta}_m$:

$$H^{-1}(\hat{\theta}_m) = H^{-1}(\theta) |_{\theta = \hat{\theta}_m} .$$

This is an asymptotic approximation, valid when $n \rightarrow \infty$. This result can be proved by expanding $p^*(\theta | x)$ in a Taylor's series about $\theta = \hat{\theta}_m$.

A sketch of the proof in the 1 dimensional case proceeds as follows.

$$\begin{aligned} p(\theta | x) &\propto p(x | \theta) \pi(\theta) \\ &= \exp\{\log[p(x | \theta) \pi(\theta)]\} \\ &= \exp\{\log[p^*(\theta | x)]\} \\ &= \exp\{l(\theta)\} \end{aligned}$$

where $l(\theta) = \log p^*(\theta | x)$. Now

$$\begin{aligned} \exp\{l(\theta)\} &\approx \exp \left\{ l(\hat{\theta}_m) + \frac{\partial}{\partial \theta} l(\theta) \Big|_{\theta=\hat{\theta}_m} (\theta - \hat{\theta}_m) + \frac{\partial^2}{\partial \theta^2} l(\theta) \Big|_{\theta=\hat{\theta}_m} \frac{(\theta - \hat{\theta}_m)^2}{2} \right\} \\ &= \exp \left\{ l(\hat{\theta}_m) - \frac{1}{2} H(\hat{\theta}_m) (\theta - \hat{\theta}_m)^2 \right\} \\ &\propto \exp \left\{ -\frac{1}{2} H(\hat{\theta}_m) (\theta - \hat{\theta}_m)^2 \right\}. \end{aligned}$$

We recognize the above as a normal density, that is,

$$N\left(\hat{\theta}_m, H(\hat{\theta}_m)^{-1}\right).$$

Note that the second term in the Taylor's expansion is 0, since

$$\frac{\partial}{\partial \theta} l(\theta) \Big|_{\theta=\hat{\theta}_m} = 0.$$

A more rigorous proof requires showing that the remainder terms in the Taylor's series become negligible as $n \rightarrow \infty$.

For the multivariate version of the proof, we have

$$\exp\{l(\theta)\} \approx \exp\left\{l(\hat{\theta}_m) - \frac{1}{2} (\theta - \hat{\theta}_m)' H(\hat{\theta}_m)(\theta - \hat{\theta}_m)\right\}.$$

Example 3.1

Suppose x_1, \dots, x_n are i.i.d. Binomial(1, θ), and suppose $\pi(\theta) = 1$, $0 < \theta < 1$. Thus

$$p(\theta | x) \propto \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}.$$

The posterior mode of θ is $\hat{\theta}_m = \frac{\sum x_i}{n}$. Also

$$p^*(\theta | x) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i},$$

$$\log p^*(\theta | x) = \sum x_i \log \theta + (n - \sum x_i) \log(1 - \theta),$$

$$\frac{\partial}{\partial \theta} \log p^*(\theta | x) = \frac{\sum x_i}{\theta} - \frac{(n - \sum x_i)}{1 - \theta},$$

$$\frac{\partial^2 \log p^*(\theta | x)}{\partial \theta^2} = \frac{-\sum x_i}{\theta^2} - \frac{(n - \sum x_i)}{(1 - \theta)^2}.$$

Thus,

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

and

$$\begin{aligned} H(\theta) \Big|_{\theta=\hat{\theta}_m} &= \frac{\Sigma x_i}{\hat{\theta}_m^2} + \frac{(n - \Sigma x_i)}{(1 - \hat{\theta}_m)^2} \\ &= \frac{n \hat{\theta}_m}{\hat{\theta}_m^2} + \frac{(n - n \hat{\theta}_m)}{(1 - \hat{\theta}_m)^2} \\ &= \frac{n}{\hat{\theta}_m} + \frac{n}{1 - \hat{\theta}_m}. \end{aligned}$$

Thus

$$\begin{aligned} (H(\hat{\theta}_m))^{-1} &= \left(\frac{n}{\hat{\theta}_m} + \frac{n}{1 - \hat{\theta}_m} \right)^{-1} \\ &= \frac{\hat{\theta}_m(1 - \hat{\theta}_m)}{n}. \end{aligned}$$

Thus, as $n \rightarrow \infty$,

$$\theta | x \rightarrow N \left(\hat{\theta}_m, \frac{\hat{\theta}_m(1 - \hat{\theta}_m)}{n} \right),$$

where $\hat{\theta}_m = \frac{1}{n} \sum_{i=1}^n x_i$.

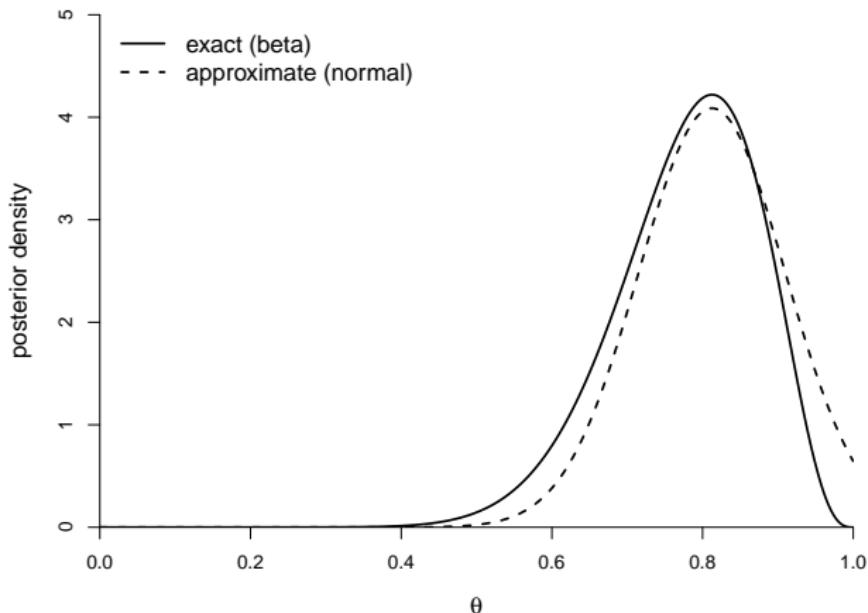


Figure 3.1: The Bayesian central limit theorem approximation of a $\text{beta}(14, 4)$, assuming $n = 16$ and $\sum x_i = 13$.

Example 3.2

Suppose x_1, \dots, x_n are i.i.d. Poisson(θ), and suppose

$$\pi(\theta) \propto \theta^{\delta_0 - 1} \exp\{-\gamma_0 \theta\}.$$

We have

$$p(\theta | x) \propto \theta^{\delta_0 + \sum x_i - 1} \exp\{-\theta(n + \gamma_0)\}$$

$$p^*(\theta | x) = \theta^{\delta_0 + \sum x_i - 1} \exp\{-\theta(n + \gamma_0)\}.$$

The posterior mode is found as

$$\begin{aligned}\frac{\partial}{\partial \theta} \log p^*(\theta | x) &= \frac{\partial}{\partial \theta} [(\delta_0 + \sum x_i - 1) \log \theta - \theta(n + \gamma_0)] \\ &= \frac{\delta_0 + \sum x_i - 1}{\theta} - (n + \gamma_0) \\ &= 0.\end{aligned}$$

Thus, $\hat{\theta}_m = \frac{\delta_0 + \sum x_i - 1}{n + \gamma_0}$.

Now

$$\begin{aligned}H(\theta) &= -\frac{\partial^2 \log p^*(\theta | x)}{\partial \theta^2} = \frac{(\delta_0 + \sum x_i - 1)}{\theta^2}, \\ H(\hat{\theta}_m) &= \frac{\delta_0 + \sum x_i - 1}{\hat{\theta}_m^2} = \frac{\delta_0 + \sum x_i - 1}{\left(\frac{\delta_0 + \sum x_i - 1}{n + \gamma_0}\right)^2} \\ &= \frac{(n + \gamma_0)^2}{\delta_0 + \sum x_i - 1}.\end{aligned}$$

Thus, as $n \rightarrow \infty$,

$$\begin{aligned}\theta | x &\rightarrow N(\hat{\theta}_m, H^{-1}(\hat{\theta}_m)) \\ &= N\left(\frac{\delta_0 + \sum x_i - 1}{n + \gamma_0}, \frac{\delta_0 + \sum x_i - 1}{(n + \gamma_0)^2}\right).\end{aligned}$$

Example 3.3

Suppose y_1, \dots, y_n are independent Binomial($1, \theta_i$), $i = 1, \dots, n$. Suppose we model the θ_i 's as functions of covariates $x'_i = (x_{i1}, \dots, x_{ip})$.

That is

$$\theta_i = \frac{\exp\{x'_i \beta\}}{1 + \exp\{x'_i \beta\}},$$

where β is a $p \times 1$ vector of regression coefficients.

The likelihood, as a function of β , is given by

$$\begin{aligned}
 p(y | \beta) &= \prod_{i=1}^n p(y_i | \beta) \\
 &= \prod_{i=1}^n \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \\
 &= \prod_{i=1}^n \left[\frac{\exp\{x_i' \beta\}}{1 + \exp\{x_i' \beta\}} \right]^{y_i} \left[\frac{1}{1 + \exp\{x_i' \beta\}} \right]^{1-y_i} \\
 &= \exp \left\{ \sum_{i=1}^n [y_i x_i' \beta - \log(1 + \exp\{x_i' \beta\})] \right\}.
 \end{aligned}$$

Suppose we specify a uniform (improper) prior for β , i.e., $\pi(\beta) \propto 1$. Then

$$p^*(\beta | y) = \exp \left\{ \sum_{i=1}^n [y_i x_i' \beta - \log(1 + \exp\{x_i' \beta\})] \right\}.$$

The posterior mode of β is the MLE of β , that is, $\hat{\beta}_m = \hat{\beta}$. This is only true since we have used a uniform prior.

Now

$$\begin{aligned}\frac{\partial}{\partial \beta_j} \log p^*(\beta | y) &= \sum_{i=1}^n \left(y_i - \frac{\exp \{x'_i \beta\}}{1 + \exp \{x'_i \beta\}} \right) x_{ij} \\ -\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log p^*(\beta | y) &= \sum_{i=1}^n x_{ij} x_{ik} \frac{\exp \{x'_i \beta\}}{(1 + \exp \{x'_i \beta\})^2}\end{aligned}$$

for $j = 1, \dots, p$ and $k = 1, \dots, p$.

In matrix notation, we have

$$\frac{\partial}{\partial \beta} \log p^*(\beta | y) = X' S ,$$

where X is the $n \times p$ matrix of covariates and

$$S = \begin{pmatrix} y_1 - \theta_1 \\ \vdots \\ y_n - \theta_n \end{pmatrix} ,$$

where

$$\theta_i = \frac{\exp \{x'_i \beta\}}{1 + \exp \{x'_i \beta\}} .$$

Also

$$\begin{aligned} H(\beta) &= - \frac{\partial^2 \log p^*(\beta | y)}{\partial \beta \partial \beta'} \\ &= X' V X , \end{aligned}$$

where V is an $n \times n$ diagonal matrix with i^{th} diagonal element

$$v_{ii} = \frac{\exp\{x_i' \beta\}}{(1 + \exp\{x_i' \beta\})^2} , \quad i = 1, \dots, n .$$

Thus, as $n \rightarrow \infty$,

$$\beta | y \rightarrow N_p(\hat{\beta}, (X' \hat{V} X)^{-1}) ,$$

where

$$\hat{V} = V \Big|_{\beta=\hat{\beta}} = \begin{bmatrix} \frac{\exp\{x_1' \hat{\beta}\}}{(1 + \exp\{x_1' \hat{\beta}\})^2} & 0 & & \\ & \ddots & & \\ 0 & & \frac{\exp\{x_n' \hat{\beta}\}}{(1 + \exp\{x_n' \hat{\beta}\})^2} & \end{bmatrix}_{n \times n} .$$

Note: The posterior distribution of β does **not** have an analytic closed form in this example.

Laplace's Method

The posterior approximation techniques given by the Bayesian central limit theorem are often referred to as first order approximations, or modal approximations. Estimates of posterior quantities may be obtained simply as the corresponding features of the approximating normal density.

These approximations may be poor if the true posterior differs greatly from normality, or if the sample size is too small for the approximation to work well.

This raises the question of whether we can obtain more accurate posterior estimates without significantly more effort in terms of higher order derivatives or complicated transformations. An affirmative answer to this question is provided by an expansion technique known as **Laplace's method**.

This technique is based on the work of Laplace (1774) and reprinted in *Statistical Science* (1986, 364-378). Tierney and Kadane (1986, JASA); Kass, Tierney, and Kadane (1988, *Bayesian Statistics 3*); and Kass, Tierney, and Kadane (1989, JASA) give an extensive discussion of Laplace's method with applications to Bayesian inference.

Laplace's method proceeds as follows:

Suppose f is a smooth function of θ , and h is a smooth function of θ with $-h$ having unique maximum $\hat{\theta}$. For simplicity, take θ to be 1 dimensional.

Suppose we wish to approximate

$$I = \int f(\theta) \exp\{-nh(\theta)\} d\theta .$$

Using Taylor's series expansions of both f and h about $\hat{\theta}$, we obtain

$$\begin{aligned} I &\approx \int \left[f(\hat{\theta}) + \frac{f'(\hat{\theta})}{1!} (\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2!} (\theta - \hat{\theta})^2 \right] \\ &\quad \times \exp \left\{ -n h(\hat{\theta}) - n h'(\hat{\theta})(\theta - \hat{\theta}) - n h''(\hat{\theta}) \frac{(\theta - \hat{\theta})^2}{2!} \right\} d\theta \\ &= \exp \left\{ -n h(\hat{\theta}) \right\} \int \left[f(\hat{\theta}) + f'(\hat{\theta}) (\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2} (\theta - \hat{\theta})^2 \right] \\ &\quad \times \exp \left\{ -\frac{(\theta - \hat{\theta})^2}{2 (n h''(\hat{\theta}))^{-1}} \right\} d\theta , \end{aligned}$$

$$\begin{aligned}
&= \exp\left\{-nh(\hat{\theta})\right\} \int \left\{ f(\hat{\theta}) \sqrt{\frac{2\pi}{nh''(\hat{\theta})}} \right\} \sqrt{\frac{nh''(\hat{\theta})}{2\pi}} \exp\left\{-\frac{(\theta - \hat{\theta})^2}{2(nh''(\hat{\theta}))^{-1}}\right\} d\theta \\
&+ \exp\left\{-nh(\hat{\theta})\right\} \int \left\{ f'(\hat{\theta})(\theta - \hat{\theta}) \sqrt{\frac{2\pi}{nh''(\hat{\theta})}} \right\} \sqrt{\frac{nh''(\hat{\theta})}{2\pi}} \exp\left\{-\frac{(\theta - \hat{\theta})^2}{2(nh''(\hat{\theta}))^{-1}}\right\} d\theta \\
&+ \exp\left\{-nh(\hat{\theta})\right\} \int \left\{ \frac{f''(\hat{\theta})}{2} (\theta - \hat{\theta})^2 \sqrt{\frac{2\pi}{nh''(\hat{\theta})}} \right\} \sqrt{\frac{nh''(\hat{\theta})}{2\pi}} \exp\left\{-\frac{(\theta - \hat{\theta})^2}{2(nh''(\hat{\theta}))^{-1}}\right\} d\theta .
\end{aligned}$$

Note that

$$\sqrt{\frac{nh''(\hat{\theta})}{2\pi}} \exp\left\{-\frac{(\theta - \hat{\theta})^2}{2(nh''(\hat{\theta}))^{-1}}\right\}$$

is the $N\left(\hat{\theta}, (n h''(\hat{\theta}))^{-1}\right)$ density, so the above equals

$$\exp\left\{-nh(\hat{\theta})\right\} f(\hat{\theta}) \sqrt{\frac{2\pi}{nh''(\hat{\theta})}} \left[1 + \frac{f'(\hat{\theta})E(\theta - \hat{\theta})}{f(\hat{\theta})} + \frac{f''(\hat{\theta})\text{Var}(\theta)}{2f(\hat{\theta})} \right] .$$

The second term inside the brackets vanishes since $E(\theta - \hat{\theta}) = 0$.

Thus we have

$$I \approx \exp \left\{ -nh(\hat{\theta}) \right\} f(\hat{\theta}) \sqrt{\frac{2\pi}{n h''(\hat{\theta})}} \left[1 + \frac{f''(\hat{\theta})}{2f(\hat{\theta})} \text{Var}(\theta) \right] ,$$

where $\text{Var}(\theta) = [n h''(\hat{\theta})]^{-1} = O(\frac{1}{n})$. Therefore, we can write

$$\begin{aligned} I &= \exp \left\{ -nh(\hat{\theta}) \right\} f(\hat{\theta}) \sqrt{\frac{2\pi}{n h''(\hat{\theta})}} [1 + O(\frac{1}{n})] \\ &\approx \exp \left\{ -n h(\hat{\theta}) \right\} f(\hat{\theta}) \sqrt{\frac{2\pi}{n h''(\hat{\theta})}} . \end{aligned}$$

If θ is p -dimensional, Laplace's method yields

$$I = \hat{I} [1 + O\left(\frac{1}{n}\right)] ,$$

where

$$\hat{I} = f(\hat{\theta}) \left(\frac{2\pi}{n} \right)^{p/2} |\tilde{\Sigma}|^{\frac{1}{2}} \exp \left\{ -nh(\hat{\theta}) \right\} ,$$

where

$$\tilde{\Sigma} = \left[\frac{\partial^2 h(\theta)}{\partial \theta_j \partial \theta_k} \right]_{\theta=\hat{\theta}}^{-1} .$$

Thus $\tilde{\Sigma}$ is a $p \times p$ matrix. Hence, \hat{I} is a **first order** approximation to I in an asymptotic sense.

Laplace's method can be used in a clever way to obtain **second order** approximations to posterior expectations of functions of θ . To this end, suppose we wish to compute the posterior expectation of a function $g(\theta)$, where we think of $-nh(\theta)$ as the unnormalized posterior density, that is,

$$-n h(\theta) = \log[p(x | \theta) \pi(\theta)] .$$

Then

$$E[g(\theta)] = \frac{\int g(\theta) \exp\{-nh(\theta)\} d\theta}{\int \exp\{-nh(\theta)\} d\theta} . \quad (3.2)$$

Here, θ may be p dimensional.

We can apply Laplace's method to the numerator using $f = g$ and for the denominator using $f = 1$. We get

$$\begin{aligned} E[g(\theta)] &= \frac{g(\hat{\theta}) \left(\frac{2\pi}{n}\right)^{p/2} |\tilde{\Sigma}|^{\frac{1}{2}} \exp\{-nh(\hat{\theta})\} [1 + O_1(\frac{1}{n})]}{\left(\frac{2\pi}{n}\right)^{p/2} |\tilde{\Sigma}|^{\frac{1}{2}} \exp\{-nh(\hat{\theta})\} [1 + O_2(\frac{1}{n})]} \\ &= g(\hat{\theta}) \left[\frac{1 + O_2(\frac{1}{n}) - O_2(\frac{1}{n}) + O_1(\frac{1}{n})}{1 + O_2(\frac{1}{n})} \right] \\ &= g(\hat{\theta}) \left[1 + \frac{O_3(\frac{1}{n})}{1 + O_2(\frac{1}{n})} \right] = g(\hat{\theta}) [1 + O(\frac{1}{n})] . \end{aligned}$$

Thus

$$\begin{aligned} E[g(\theta)] &= g(\hat{\theta}) [1 + O(\frac{1}{n})] \\ &\approx g(\hat{\theta}) . \end{aligned}$$

Therefore, $g(\hat{\theta})$ is a **first order** approximation to the posterior mean of $g(\theta)$. We see that this naive application of Laplace's method produces the same estimator (with the same order of accuracy) as the modal approximation.

We can get a **second order** approximation by introducing the following trick (Tierney and Kadane, 1986, JASA). Suppose $g(\theta) > 0$ for all θ . We can write the numerator of $E(g(\theta))$ (i.e., Equation (3.2)), as

$$\exp\{\log(g(\theta)) - n h(\theta)\} \equiv \exp\{-n h^*(\theta)\},$$

where $h^*(\theta) = -\frac{1}{n} \log g(\theta) + h(\theta)$. Now we apply Laplace's method with $f = 1$ in **both** the numerator and denominator. The result is

$$E[g(\theta)] = \frac{|\Sigma^*|^{\frac{1}{2}} \exp\{-n h^*(\theta^*)\}}{|\tilde{\Sigma}|^{\frac{1}{2}} \exp\{-n h(\hat{\theta})\}} [1 + O\left(\frac{1}{n^2}\right)], \quad (3.3)$$

where θ^* is the maximum of $-h^*$ and

$$\Sigma^* = \left[\frac{\partial^2 h^*(\theta)}{\partial \theta_j \partial \theta_k} \right]_{\theta=\theta^*}^{-1}.$$

This clever application of Laplace's method thus provides a **second order** approximation to the posterior mean of $g(\theta)$, but still requires only the calculation of first and second derivatives of the log-posterior. By contrast, most other second order expansion methods require computation of third and perhaps higher order derivatives. The improvement in accuracy in (3.3) above comes since the leading terms in the two errors (both $O(\frac{1}{n})$) are **identical**, and thus **cancel** when the ratio is taken (see Tierney and Kadane, 1986, JASA).

For functions $g(\theta)$ which are not strictly positive, a simple solution is to add a large constant c to $g(\theta)$, apply Laplace's method (3) to this function, and then subtract c out in the final answer. Alternatively, Tierney, Kass, and Kadane (1989, JASA) recommend applying Laplace's method to approximate the moment generating function $E[\exp\{sg(\theta)\}]$, whose integrand is always positive, and then differentiating the result after approximation to obtain the final answer. Note that

$$\psi(s) = E[\exp\{sg(\theta)\}]$$

is the MGF of $g(\theta)$. Tierney, Kass, and Kadane (1989) show that these two approaches are asymptotically equivalent as $c \rightarrow \infty$.

Improved approximation of marginal densities is also possible using Laplace's method. Suppose $\Theta = \Theta_1 \times \Theta_2$, and we want an estimate of

$$p(\theta_1 | x) = c_1 \int \exp\{\tilde{p}(\theta_1, \theta_2 | x)\} d\theta_2 ,$$

where $\tilde{p}(\theta_1, \theta_2 | x) = \log[p(x | \theta) \pi(\theta)]$ and c_1 is an unknown normalizing constant. Taking $f = 1$ and $h = -n^{-1} \tilde{p}$, we may use the Laplace approximation in (3.3) again to obtain

$$\hat{p}(\theta_1 | x) = c^* |\tilde{\Sigma}(\theta_1)|^{\frac{1}{2}} \exp\{\tilde{p}(\theta_1, \hat{\theta}_2(\theta_1) | x)\} ,$$

where c^* is a **new** normalizing constant, $\hat{\theta}_2(\theta_1)$ is the maximum of $\tilde{p}(\theta_1, \cdot)$, and

$$\tilde{\Sigma}(\theta_1) = -n \left[\frac{\partial^2 \tilde{p}(\theta_1, \cdot)}{\partial \theta_{2j} \partial \theta_{2k}} \right]^{-1} \Big|_{\theta_2=\hat{\theta}_2(\theta_1)} .$$

In order to plot this density, the formula for $\hat{p}(\theta_1 | x)$ above would need to be evaluated over a grid of θ_1 values, after which the normalizing constant c^* could be estimated. This approximation has relative error of order $O(n^{-3/2})$ on neighborhoods about the mode that shrink at a rate $n^{-\frac{1}{2}}$.

There are several advantages of Laplace's method.

- ▶ It is a computationally quick procedure, since it does not require an iterative algorithm.
- ▶ It replaces numerical integration with numerical differentiation, which is often easier and more stable numerically.
- ▶ It is a deterministic algorithm (i.e., does not rely on random numbers), so two different analysts should be able to produce a common answer given the same dataset, model, and prior distribution.
- ▶ It greatly reduces the computational complexity in any study of robustness, i.e., an investigation of how sensitive our conclusions are to modest changes in the prior or likelihood function.

For example, suppose we wish to find the posterior expectation of a function of interest $g(\theta)$ under a **new** prior distribution, $\pi_{\text{new}}(\theta)$.

We can write

$$E_{\text{new}}[g(\theta) \mid x] = \frac{\int g(\theta) p(x \mid \theta) \pi(\theta) b(\theta) d\theta}{\int p(x \mid \theta) \pi(\theta) b(\theta) d\theta}$$

where $b(\cdot)$ is the **perturbation function**, which in this case is $b(\theta) = \frac{\pi_{\text{new}}(\theta)}{\pi(\theta)}$. We can avoid starting the posterior calculations from scratch by using a result due to Kass, Tierney and Kadane (1989, Biometrika), namely

$$E_{\text{new}}[g(\theta) \mid x] \approx \frac{b(\theta^*)}{b(\hat{\theta})} E[g(\theta) \mid x],$$

where θ^* and $\hat{\theta}$ maximize $\log[g(\theta) p(x \mid \theta) \pi(\theta)]$ and $\log[p(x \mid \theta) \pi(\theta)]$, respectively.

Note that if equation (3.3) is used to compute the original posterior expectation $E[g(\theta)|x]$, then θ^* and $\hat{\theta}$ are already available.

Hence, any number of alternate priors may be investigated simply by evaluating the ratios $\frac{b(\theta^*)}{b(\hat{\theta})}$ and no new derivative calculations are required.

The Laplace method also has several limitations.

- ▶ For the approximation to be valid, the posterior distribution must be unimodal, or nearly so.
- ▶ The accuracy depends on the parameterization used, for example, θ or $\log(\theta)$, and the correct one may be difficult to ascertain.
- ▶ The size of the data set, n , must be fairly large.
- ▶ Moreover it is hard to judge how large is “large enough.” That is, the second order accuracy provided by equation (3.3) is comforting and surely better than the first order accuracy provided by the normal approximation, but for any given dataset, we will still lack a numerical measurement of how far our approximate posterior expectations are from their exact values.

Worse yet, since the asymptotics are in n , the size of the dataset, we will not be able to improve the accuracy of our approximations without collecting additional data!

- ▶ For moderate to high-dimensional θ (say bigger than 10), Laplace's method will rarely be of sufficient accuracy, and numerical computation of the associated Hessian matrices will be prohibitively difficult.

- ▶ Unfortunately, high-dimensional models (such as random effects models), which feature a parameter for every subject in the study, are fast becoming the norm in many branches of applied statistics, especially those involving biostatistical applications.

For these reasons, practitioners are increasingly turning to iterative methods, especially those involving Monte Carlo sampling for posterior calculation.

Example 3.4

Suppose x_1, \dots, x_n are i.i.d. Binomial($1, \theta$) with $\pi(\theta) = 1$, $0 < \theta < 1$.

Using Laplace's method, we want to compute $E(\theta|x)$.

We know that an **exact** expression for $E(\theta|x)$ is given by

$$E(\theta|x) = \frac{1 + \sum x_i}{n + 2} = \frac{1 + n\bar{x}}{n + 2}.$$

Let us apply Laplace's methods using the results of Tierney and Kadane (1986, JASA).

We have

$$\begin{aligned}
 E(\theta|x) &= \frac{\int_0^1 \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} d\theta}{\int_0^1 \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} d\theta} \\
 &= \frac{\int_0^1 \exp\{\log(\theta)\} \exp\{-n[-\bar{x}\log\theta - (1-\bar{x})\log(1-\theta)]\} d\theta}{\int_0^1 \exp\{-n[-\bar{x}\log\theta - (1-\bar{x})\log(1-\theta)]\} d\theta} \\
 &= \frac{\int_0^1 \exp\left\{-n\left[-\frac{1}{n}\log(\theta) - \bar{x}\log(\theta) - (1-\bar{x})\log(1-\theta)\right]\right\} d\theta}{\int_0^1 \exp\{-n[-\bar{x}\log(\theta) - (1-\bar{x})\log(1-\theta)]\} d\theta}.
 \end{aligned}$$

Thus we have

$$h^*(\theta) = -\frac{1}{n} \log \theta - \bar{x} \log(\theta) - (1 - \bar{x}) \log(1 - \theta) ,$$

$$h(\theta) = -\bar{x} \log(\theta) - (1 - \bar{x}) \log(1 - \theta) .$$

Now we need to find the maxima of $-h^*$ and $-h$.

$$-h^*(\theta) = \frac{1}{n} \log \theta + \bar{x} \log(\theta) + (1 - \bar{x}) \log(1 - \theta)$$

$$\frac{\partial}{\partial \theta} (-h^*(\theta)) = \frac{1}{n\theta} + \frac{\bar{x}}{\theta} - \frac{(1-\bar{x})}{1-\theta} = 0$$

$$= \frac{1+n\bar{x}}{n\theta} - \frac{(1-\bar{x})}{1-\theta} = 0$$

$$\implies \frac{(1-\theta)(1+n\bar{x}) - (1-\bar{x})n\theta}{n\theta(1-\theta)} = 0$$

$$\implies 1 - \theta + n\bar{x} - \theta n\bar{x} - n\theta + n\theta\bar{x} = 0$$

$$1 + n\bar{x} = \theta + n\theta ,$$

Thus,

$$\theta^* = \frac{1+n\bar{x}}{n+1} .$$

Similarly, the maximum of $-h(\theta)$ is

$$\frac{\partial}{\partial \theta}[-h(\theta)] = \frac{\bar{x}}{\theta} - \frac{(1-\bar{x})}{1-\theta} = 0,$$

leading to

$$\hat{\theta} = \bar{x} .$$

Now we need to compute $\tilde{\Sigma}$ and Σ^* .

$$\frac{\partial^2 h^*(\theta)}{\partial \theta^2} = \frac{1}{n\theta^2} + \frac{\bar{x}}{\theta^2} + \frac{(1-\bar{x})}{(1-\theta)^2} = \frac{(1+n\bar{x})}{n\theta^2} + \frac{(1-\bar{x})}{(1-\theta)^2} ,$$

and thus

$$\begin{aligned}\Sigma^* &= \left(\frac{\partial^2 h^*(\theta)}{\partial \theta^2} \right)^{-1} \Big|_{\theta=\theta^*} = \left(\frac{\frac{1+n\bar{x}}{n} \left[\frac{1+n\bar{x}}{n+1} \right]^2}{\left(1 - \frac{(1+n\bar{x})}{n+1} \right)^2} + \frac{(1-\bar{x})}{\left(1 - \frac{(1+n\bar{x})}{n+1} \right)^2} \right)^{-1} \\ &= \left(\frac{(n+1)^2}{n(1+n\bar{x})} + \frac{(n+1)^2}{n^2(1-\bar{x})} \right)^{-1}.\end{aligned}$$

$$\begin{aligned}\tilde{\Sigma} &= \left(\frac{\partial^2 h(\theta)}{\partial \theta^2} \right)^{-1} \Big|_{\theta=\hat{\theta}} = \left(\frac{\bar{x}}{\hat{\theta}^2} + \frac{(1-\bar{x})}{(1-\hat{\theta})^2} \Big|_{\theta=\hat{\theta}} \right)^{-1} \\ &= \left(\frac{\bar{x}}{\bar{x}^2} + \frac{(1-\bar{x})}{(1-\bar{x})^2} \right)^{-1} \\ &= \left(\frac{1}{\bar{x}} + \frac{1}{1-\bar{x}} \right)^{-1} \\ &= \bar{x}(1-\bar{x}).\end{aligned}$$

Now the Laplace approximation is given by

$$E[g(\theta)|x] \approx \frac{|\Sigma^*|^{\frac{1}{2}} \exp \{-nh^*(\theta^*)\}}{|\tilde{\Sigma}|^{\frac{1}{2}} \exp \{-nh(\hat{\theta})\}}, \quad g(\theta) = \theta.$$

Thus, we get

$$E[\theta|x] \approx \frac{\left[\frac{(n+1)^2}{n(1+n\bar{x})} + \frac{(n+1)^2}{n^2(1-\bar{x})} \right]^{-\frac{1}{2}} \exp \left\{ \log \left(\frac{1+n\bar{x}}{n+1} \right) + n\bar{x} \log \left(\frac{1+n\bar{x}}{n+1} \right) + A \right\}}{(\bar{x}(1-\bar{x}))^{\frac{1}{2}} \exp \{ n\bar{x} \log(\bar{x}) + n(1-\bar{x}) \log(1-\bar{x}) \}}$$

where $A = n(1-\bar{x}) \log \left(1 - \frac{1+n\bar{x}}{n+1} \right)$.

Exercise 3.1

- (a) Using Laplace's method, compute $E(\exp\{\theta\}|x)$.
- (b) Compute the Laplace approximation for $E[\exp\{\theta\}|x]$ using $n = 5, 15, 30, 100$ and $\bar{x} = .4$.
- (c) Find the Laplace approximation for the normalizing constant of the posterior distribution, and compare it with the exact result for $n = 5, 15, 30, 100$ and $\bar{x} = .4$.

Example 3.5

Using the result from Example 3.4, compute the Laplace approximation for $n = 5, 15, 30, 100$ with $\bar{x} = .4$. Compare with the exact answer.

R code:

```

n <- c(5,15,30,100)
xbar <- .4

thetastar <- (1 + n * xbar) / (n + 1)
thetahat <- xbar

sigmastar <- ( (1 + n * xbar) / (n * thetastar^2) + (1 - xbar) / (1 - thetastar)^2 ) ^(-1)
sigmatilde <- xbar * (1 - xbar)

hstar <- (-1/n) * log(thetastar) - xbar*log(thetastar) - (1 - xbar) * log(1 - thetastar)
h <- -1.0 * xbar * log(thetahat) - (1 - xbar) * log(1 - thetahat)

laplace <- (sigmastar^.5 * exp(-n * hstar)) / (sigmatilde^.5 * exp(-n*h))
exact <- (1 + n*xbar) / (n + 2)

> laplace
[1] 0.4212307 0.4108263 0.4060076 0.4019385

> exact
[1] 0.4285714 0.4117647 0.4062500 0.4019608

```

Noniterative Monte Carlo Methods

1. Direct Sampling

Approximations to integrals can be carried out by Monte Carlo integration. The definition of Monte Carlo integration is as follows. Suppose a random variable θ has density generically denoted by $p(\theta)$, and we wish to compute $E[f(\theta)]$.

This is given by

$$\gamma = E[f(\theta)] = \int f(\theta) p(\theta) d\theta.$$

Then if $\theta_1, \dots, \theta_N$ are i.i.d. samples from $p(\theta)$, we have

$$\hat{\gamma} = \frac{1}{N} \sum_{j=1}^N f(\theta_j), \tag{3.4}$$

which converges to $E[f(\theta)]$ with probability 1 as $N \rightarrow \infty$, by the strong law of large numbers.

In Bayesian inference, $p(\theta)$ typically corresponds to the posterior distribution of θ , $p(\theta|x)$, and thus $E[f(\theta)]$ is the posterior mean of $f(\theta)$.

Hence the computation of posterior expectations requires only a sample of size N from the posterior distribution.

Thus, we must be able to directly sample from the posterior distribution in order to use the Monte Carlo approximation. Notice that as N increases, the quality of the approximation increases. The Monte Carlo size, N , is within our control, whereas the sample size n is not.

Another constraint with asymptotic methods is that the structure of (3.4) also allows us to evaluate its accuracy for any fixed N . Since $\hat{\gamma}$ is itself a sample mean of independent observations, we have

$$\text{Var}(\hat{\gamma}) = \frac{1}{N} \text{Var}(f(\theta)).$$

But $\text{Var}(f(\theta))$ can be estimated by the sample variance of the $f(\theta_j)$ values, so that a standard error estimate of $\hat{\gamma}$ is given by

$$\text{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)} \sum_{j=1}^N (f(\theta_j) - \hat{\gamma})^2}.$$

Finally, the CLT implies that $\hat{\gamma} \pm 2 \text{ se}(\hat{\gamma})$ provides an approximate 95% interval for the true value of the posterior mean γ .

Remark 3.1

While it may seem strange for us to recommend use of a frequentist interval estimation here, Monte Carlo simulations provide one (and perhaps only one) example where they are clearly appropriate!

In addition, note that we can also estimate quantities such as

$$p = P(a < f(\theta) < b|x)$$

using Monte Carlo methods.

An estimate of p is simply

$$\hat{p} = \frac{\text{number of sample } f(\theta_j)\text{'s } \in (a, b)}{N}.$$

The associated simulation standard error estimate is

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{N}}.$$

This suggests that a histogram of the sampled θ_j 's would estimate the posterior itself, since the probability in each histogram bin converges to the true bin probability.

Alternatively, we could use a **kernel density** estimate to “smooth” the histogram.

$$\hat{p}(\theta|x) = \frac{1}{Nk_N} \sum_{j=1}^N K\left(\frac{\theta - \theta_j}{k_N}\right),$$

where K is a “kernel” density (typically a normal or rectangular distribution) and k_N is a **window width** satisfying $k_N \rightarrow 0$ and $Nk_N \rightarrow \infty$ as $N \rightarrow \infty$.

Details of density estimation are given in a book by Silverman (1986, Chapman and Hall).

Example 3.6

Suppose x_1, \dots, x_n are i.i.d. $N(\mu, \sigma^2)$, $\tau = \frac{1}{\sigma^2}$, and $\pi(\mu, \tau) \propto \tau^{-1}$.

Recall that

$$\begin{aligned} \mu|\tau, x &\sim N(\bar{x}, \frac{1}{n\tau}), \\ \text{and } \tau|x &\sim \text{gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right), \end{aligned}$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

We can generate samples from the joint posterior of (μ, τ) as follows:

1. Sample $\tau_j \sim \text{gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$.
2. Then sample $\mu_j \sim N(\bar{x}, \frac{1}{n\tau_j})$, $j = 1, \dots, N$.

This process then creates the set $\{(\mu_j, \tau_j), j = 1, \dots, N\}$ from $p(\mu, \tau|x)$. To estimate the posterior mean of μ , we would use

$$\hat{E}(\mu|x) = \frac{1}{N} \sum_{j=1}^N \mu_j .$$

To obtain a 95% HPD interval for μ , we simply use the empirical .025 and .975 quantiles of the sample of μ_j values.

Estimates of functions of the parameters are also easily obtained. For example, suppose we want to estimate the posterior distribution of $\gamma = \frac{\sigma}{\mu} = \frac{\tau^{-\frac{1}{2}}}{\mu}$, which is known as the **coefficient of variation**.

We simply define the transformed Monte Carlo samples

$\gamma_j = \frac{\tau_j^{-\frac{1}{2}}}{\mu_j}$, $j = 1, \dots, N$, and create a histogram or kernel density estimate based on these values.

As a final illustration, suppose we wanted to estimate

$$P(z > c|x) = \int_{\Theta} \left[\int_c^{\infty} p(z|\theta) dz \right] p(\theta|x) d\theta ,$$

where z is a **future observation**. Note that the inner integral can be written as

$$P(z > c|\theta) = 1 - \Phi\left(\frac{c - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{c - \mu}{\tau^{-\frac{1}{2}}}\right) ,$$

where $\Phi(\cdot)$ is the standard normal c.d.f.

Thus

$$\begin{aligned} P(z > c|x) &= E_{\theta|x} \left[1 - \Phi\left(\frac{c - \mu}{\tau^{-\frac{1}{2}}}\right) \right] \\ &\approx \frac{1}{N} \sum_{j=1}^N \left[1 - \Phi\left(\frac{c - \mu_j}{\tau_j^{-\frac{1}{2}}}\right) \right] . \end{aligned}$$

2. Indirect Methods

If we cannot **directly sample** from the joint posterior, as in the previous example, then we cannot use **direct Monte Carlo**. We need to search for ways to sample from the joint posterior **indirectly**.

Several methods have been proposed in the literature for **indirect sampling**. We will discuss three of them. These are

- i) importance sampling
- ii) rejection sampling
- iii) the weighted bootstrap

Importance Sampling

This approach was outlined carefully by Hammersley and Handscombe (1964, Monte Carlo Methods, Chapman and Hall) and championed for Bayesian analysis by Geweke (1989, *Econometrica*).

Suppose we wish to approximate a posterior expectation, say

$$E[f(\theta)|x] = \frac{\int f(\theta) L(\theta) \pi(\theta) d\theta}{\int L(\theta) \pi(\theta) d\theta} ,$$

where $L(\theta) \propto p(x|\theta)$ is the likelihood function of θ .

Suppose we can roughly approximate the normalized posterior, $\frac{L(\theta) \pi(\theta)}{c}$, by some density $g(\theta)$, from which we can easily sample, say a multivariate t density or perhaps some variation of a multivariate t .

We call $g(\theta)$ the **importance function**.

Then defining the weight function

$$w(\theta) = \frac{L(\theta) \pi(\theta)}{g(\theta)} ,$$

we have

$$\begin{aligned} E[f(\theta)|x] &= \frac{\int f(\theta) w(\theta) g(\theta) d\theta}{\int w(\theta) g(\theta) d\theta} \\ &\approx \frac{\frac{1}{N} \sum_{j=1}^N f(\theta_j) w(\theta_j)}{\frac{1}{N} \sum_{j=1}^N w(\theta_j)} \end{aligned} \tag{3.5}$$

where the θ_j are i.i.d. samples from $g(\theta)$.

The importance sampling algorithm is given by

- ▶ Draw $\theta_1, \dots, \theta_N \stackrel{iid}{\sim} g(\theta)$
- ▶ Compute $\tilde{E}(f(\theta)|x) = \frac{\sum_{j=1}^n f(\theta_j) w(\theta_j)}{\sum_{j=1}^n w(\theta_j)} .$

Note that we do not get samples from $\frac{L(\theta)\pi(\theta)}{c}$ as a result of this procedure.

How closely $g(\theta)$ resembles $\frac{L(\theta) \pi(\theta)}{c}$ (i.e., the normalized posterior) controls how good the approximation in (3.5) is. To see this, note that if $g(\theta)$ is a good approximation, the weights will all be roughly equal, which in turn will minimize the variance of the numerator and denominator of (3.5).

If $g(\theta)$ is a poor approximation, many of the weights will be close to **zero**, and thus few θ_j 's will dominate the sums, producing an inaccurate approximation.

Example 3.7

Suppose $g(\theta)$ is taken to be the relatively light tailed normal distribution, but that $\frac{L(\theta) \pi(\theta)}{c}$ has much heavier Cauchy-like tails. Then it will take many

draws from g to obtain a few samples in these tails, and these points will have disproportionately large weights (since g will be small relative to $\frac{L(\theta) \pi(\theta)}{c}$ for these points), thus destabilizing the estimate in (3.5).

As a result, a very large N will be required to obtain an accurate approximation.

We may check the accuracy of the approximation in (3.5) using the following formula:

$$\text{Var}\left(\frac{\bar{z}}{\bar{y}}\right) \approx \frac{1}{N} \left(\frac{\hat{\sigma}_z^2}{\bar{y}^2} + \frac{\bar{z}^2 \hat{\sigma}_y^2}{\bar{y}^4} - \frac{2\bar{z}\hat{\sigma}_{zy}}{\bar{y}^3} \right),$$

where

$$\hat{\sigma}_z^2 = \frac{1}{N-1} \sum_{j=1}^N (z_j - \bar{z})^2, \quad \hat{\sigma}_y^2 = \frac{1}{N-1} \sum_{j=1}^n (y_j - \bar{y})^2$$

and

$$\hat{\sigma}_{zy} = \frac{1}{N-1} \sum_{j=1}^N (z_j - \bar{z})(y_j - \bar{y}).$$

Here we have

$$\begin{aligned} z_j &= f(\theta_j) w(\theta_j), \\ y_j &= w(\theta_j). \end{aligned}$$

Derivation of formula

Suppose z and y are random variables. Let $f(z, y) = \frac{z}{y}$.

We expand $f(z, y)$ in a Taylor series about $(E(z), E(y)) = (\mu_x, \mu_y)$.

$$f(z, y) \approx f(\mu_z, \mu_y) + (z - \mu_z) \left. \frac{\partial f(z, y)}{\partial z} \right|_{(\mu_z, \mu_y)}$$

$$+ (y - \mu_y) \left. \frac{\partial f(z, y)}{\partial y} \right|_{(\mu_z, \mu_y)}$$

$$\left. \frac{\partial f(z, y)}{\partial z} \right|_{(\mu_z, \mu_y)} = \left. \frac{1}{y} \right|_{(\mu_z, \mu_y)} = \frac{1}{\mu_y}$$

$$\left. \frac{\partial f(z, y)}{\partial y} \right|_{(\mu_z, \mu_y)} = -\left. \frac{z}{y^2} \right|_{(\mu_z, \mu_y)} = -\frac{\mu_z}{\mu_y^2}$$

Thus

$$\text{Var}(f(z, y)) = \text{Var}\left(\frac{z}{y}\right) \approx \text{Var}(z) \left(\frac{1}{\mu_y}\right)^2 + \text{Var}(y) \frac{\mu_z^2}{\mu_y^4} - 2 \frac{\mu_z}{\mu_y^3} \text{Cov}(z, y)$$

Now let $z = \bar{z}$, $y = \bar{y}$, $\bar{z} = \frac{1}{N} \sum_{i=1}^N z_i$, $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$. Then

$$\begin{aligned}\text{Var}\left(\frac{\bar{z}}{\bar{y}}\right) &= \frac{\text{Var}(\bar{z})}{\mu_y^2} + \text{Var}(\bar{y}) \frac{\mu_z^2}{\mu_y^4} - 2 \frac{\mu_z}{\mu_y^3} \text{Cov}(\bar{z}, \bar{y}) \\ &= \frac{\sigma_z^2}{N \mu_y^2} + \frac{\sigma_y^2 \mu_z^2}{N \mu_y^4} - \frac{2 \mu_z \sigma_{zy}}{N \mu_y^3} \\ &= \frac{1}{N} \left[\frac{\sigma_z^2}{\mu_y^2} + \frac{\sigma_y^2 \mu_z^2}{\mu_y^4} - \frac{2 \mu_z \sigma_{zy}}{\mu_y^3} \right].\end{aligned}$$

Thus

$$\text{Var}\left(\frac{\bar{z}}{\bar{y}}\right) = \frac{1}{N} \left[\frac{\hat{\sigma}_z^2}{\hat{\mu}_y^2} + \frac{\hat{\sigma}_y^2 \hat{\mu}_z^2}{\hat{\mu}_y^4} - \frac{2 \hat{\mu}_z \hat{\sigma}_{zy}}{\hat{\mu}_y^3} \right],$$

where

$$\hat{\sigma}_z^2 = \frac{1}{N-1} \sum_{i=1}^N (z_i - \bar{z})^2$$

$$\hat{\mu}_y = \frac{1}{N} \sum_{i=1}^N y_i = \bar{y}$$

$$\hat{\sigma}_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2$$

$$\hat{\mu}_z = \frac{1}{N} \sum_{i=1}^N z_i = \bar{z}$$

Example 3.8

Suppose we wish to sample values from a t_{15} distribution. (You can think of this as your posterior distribution if you like). Suppose that we cannot sample values from this distribution directly but that we can sample values from a $g = t_2$ distribution.

First, we take a look at the two densities using the following R code to generate Figure 3.2:

```
# Set possible theta values
theta.grid <- seq(-10, 10, 0.1)

# Create a blank graph to hold the densities later
plot( theta.grid, dt(theta.grid, df = 2), type = "n", cex.lab = 1.25,
      xlab = expression(theta), ylab = "", ylim = c(-0.02,0.5) )

# Add the t_2 importance function with dotted line
lines( theta.grid, dt(theta.grid, df = 2), lty = 2 )

# Add the t_15, which is what really interests us
lines( theta.grid, dt(theta.grid, df = 15), lty = 1 )

# Add legend
legend( "topright", c("t(15)","t(2)", "t(2)"), lty = c(1,2), bty = "n", cex = 1.25 )
```

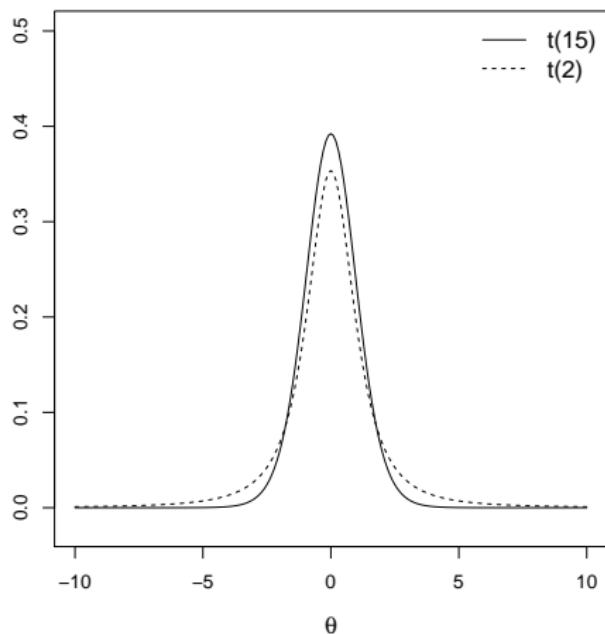


Figure 3.2:

Next, we draw 1000 samples from the t_2 distribution and plot them.

```
N <- 1000
set.seed(779)
thetas <- rt(N, df = 2)
points( thetas, rep(-.01, N), pch = 10 ) # plot points on theta axis
```

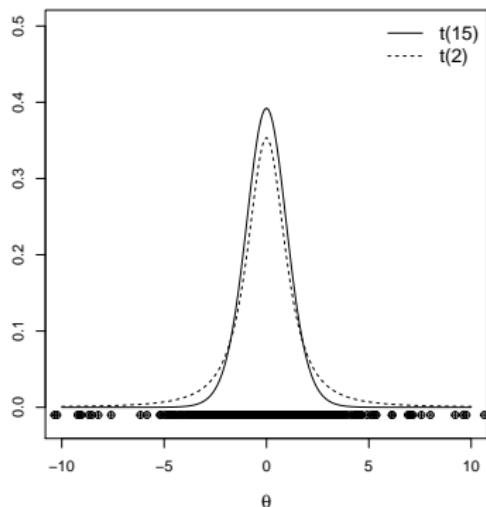


Figure 3.3:

Now we compute the importance sampling weights and then calculate the posterior mean of θ .

```
# Compute the importance sampling weights
w.theta <- dt(thetas, df = 15) * 1 / dt(thetas, df = 2)

# To compute posterior mean of theta (first moment), set f1 = theta
f1.theta <- thetas

# Estimate of posterior E[theta^2]
numerator <- sum(f1.theta * w.theta) / N
denominator <- sum(w.theta) / N
post.mean.theta <- numerator / denominator

> post.mean.theta
[1] -0.003238423

> summary(w.theta)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0000 1.0798 1.1308 0.9919 1.1848 1.2165
```

The estimate of the posterior mean of θ is -0.0032, and the interquartile range of the weights is (1.080, 1.185).

Next, we calculate the posterior variance and a confidence interval around our estimate.

```
# Compute the posterior variance of theta
# To compute 2nd moment, set f2 = theta^2
f2.theta <- thetas^2

# Importance sampling weights and denominator stay the same
numerator <- sum(f2.theta * w.theta) / N

# Estimate of posterior E[theta^2]
post.mean.theta2 <- numerator / denominator
cat(post.mean.theta2, "\n")

# Estimate of posterior Var(theta) = E(theta^2) - E(theta)^2
post.var.theta <- post.mean.theta2 - (post.mean.theta)^2

> post.var.theta
[1] 1.136864

> post.mean.theta + c(-1,1) * 1.96 * sqrt(post.var.theta)
[1] -2.093066 2.086589
```

The estimate of the posterior variance of θ is 1.137, and the 95% confidence interval for the posterior mean of θ is (-2.093, 2.087).

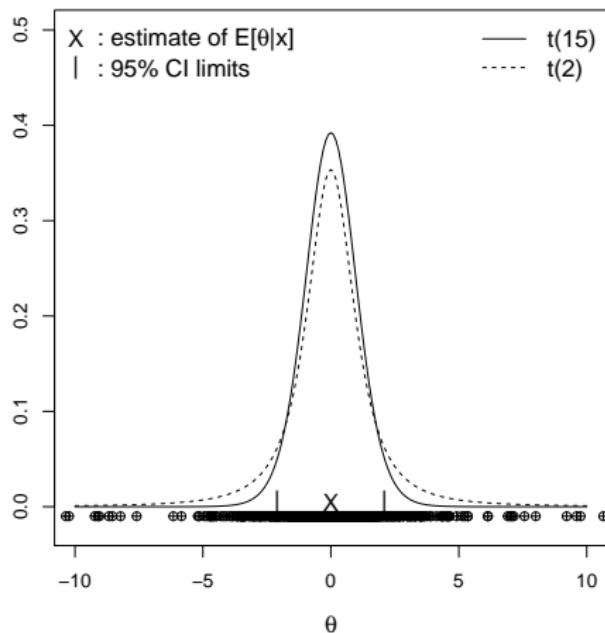


Figure 3.4:

Now suppose that we're using a "bad" importance function, for example a uniform density. Below is the code used to re-do the calculations with a uniform importance function. What, if any, was the advantage of using the "better" importance function?

```
# Sample from g
N <- 1000
set.seed(779)
thetas <- runif(N, -10, 10)
points( thetas, rep(-.01, N), pch = 10 )

# Compute the importance sampling weights
w.theta <- dt(thetas, df = 15) * 1 / (1 / 20)

# Estimate of posterior E[theta]
f1.thetas <- thetas
numerator <- sum(f1.thetas * w.theta) / N
denominator <- sum(w.theta) / N
post.mean.theta <- numerator / denominator

# Estimate of posterior E[theta^2]
f2.theta <- thetas^2
numerator <- sum(f2.theta * w.theta) / N
post.mean.theta2 <- numerator / denominator

# Estimate of posterior Var(theta) = E(theta^2) - E(theta)^2
post.var.theta <- post.mean.theta2 - (post.mean.theta)^2

> post.mean.theta
[1] 0.204918

> summary(w.theta)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.000001 0.000021 0.003784 1.071545 0.733187 7.846710
```

Here, the estimate of the posterior mean of θ is 0.2049, and the interquartile range of the weights is given by (0.00002, 1.73319).

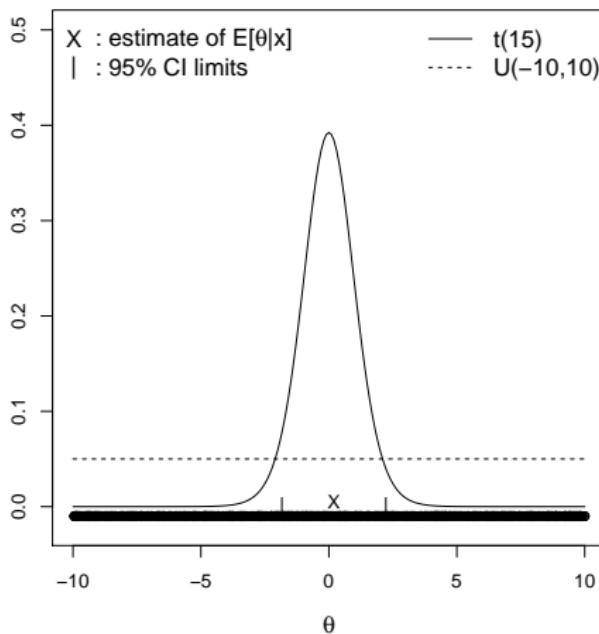


Figure 3.5: bad choice of g

Rejection Sampling

Rejection sampling is an extremely general and quite common method of random variate generation. Excellent summaries are given in the books by Ripley (1987, John Wiley and Sons) and Devroye (1986, Springer-Verlag). In this method, instead of trying to approximate the normalized posterior

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

we try to “blanket” it.

That is, suppose there exists an identifiable constant $M > 0$ and a **smooth density** $g(\theta)$, called the **envelope function**, such that

$$L(\theta) \pi(\theta) < Mg(\theta)$$

for all θ .

The **rejection method** proceeds as follows:

- i) Generate $\theta_j \sim g(\theta)$,
- ii) Generate $U \sim \text{Uniform}(0, 1)$,
- iii) If $MUg(\theta_j) < L(\theta_j) \pi(\theta_j)$, then **accept** θ_j ; otherwise **reject** θ_j ,
- iv) Return to step i) and repeat until the desired sample $\{\theta_j, j = 1, \dots, N\}$ is obtained. The members of this sample will then be random variables from $p(\theta|x)$.

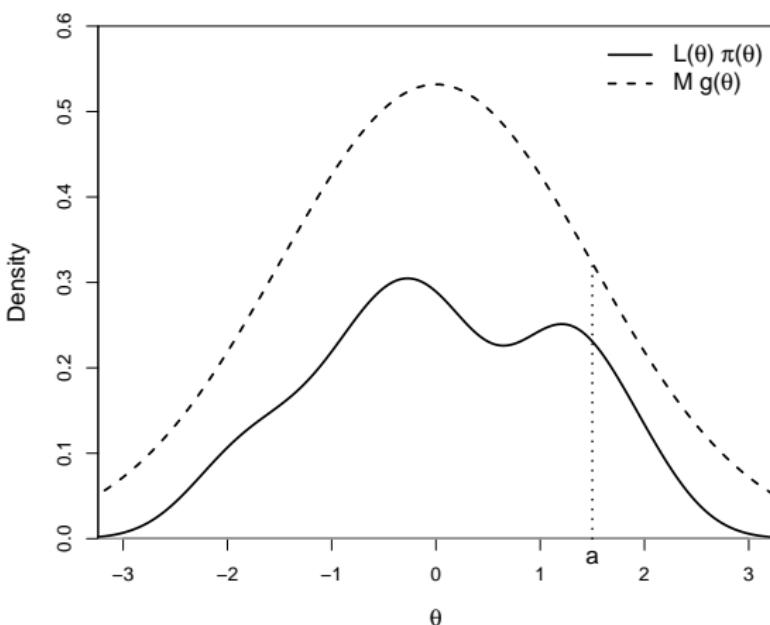


Figure 3.6: proper rejection envelope

Heuristic Justification of Rejection Algorithm

Consider a fairly large sample of points generated from $g(\theta)$. An appropriately scaled histogram of these points would have roughly the same shape as the curve labeled $Mg(\theta)$ in Figure 3.6.

Now consider the histogram bar centered at the point labeled a in Figure 3.6. The rejection step in the algorithm has the effect of slicing off the top portion of the bar (i.e., the portion between the two curves), since only those points having $MUg(\theta_j)$ values below the lower curve are retained.

But this is true for every potential value of a along the horizontal axis, so a histogram of the accepted θ_j values would mimic the shape of the lower curve, which of course is proportional to the posterior $p(\theta|x)$, as desired.

Intuition suggests that M should be chosen as small as possible, so as not to waste samples unnecessarily. This is easy to confirm, since if k denotes the number of iterations required to get one acceptable candidate θ_j , then k is a geometric random variable.

That is,

$$P(k = i) = (1 - p)^{i-1} p ,$$

where p is the probability of acceptance. So $P(k = i)$ decreases monotonically and at an exponential rate. It can be shown that

$$p = \frac{c}{M} ,$$

where

$$c = \int L(\theta) \pi(\theta) d\theta .$$

Since $E(k) = p^{-1} = \frac{M}{c}$, we do indeed want to minimize M .

Note that the geometric distribution has mean $E(K) = \frac{1}{p} = \frac{M}{c}$, and thus we do indeed want to minimize M . Note that if $p(\theta | x)$ were available as the g function, we would choose the minimal acceptable value $M = c$, obtaining an acceptance probability of 1.

Like an importance sampling density, the envelope density g should be similar to the posterior in general appearance, but with heavier tails and sharper infinite peaks in order to assure that there are sufficiently many rejection candidates available across its entire domain.

One also has to be careful that Mg is actually an “envelope” for the unnormalized posterior $L(\theta)\pi(\theta)$. To see what happens if this condition is not met, suppose

$$S_M = \{\theta : L(\theta)\pi(\theta) > Mg(\theta)\} .$$

Figure 3.7 illustrates an example of a deficient rejection envelope with $S_M = (a, b)$.

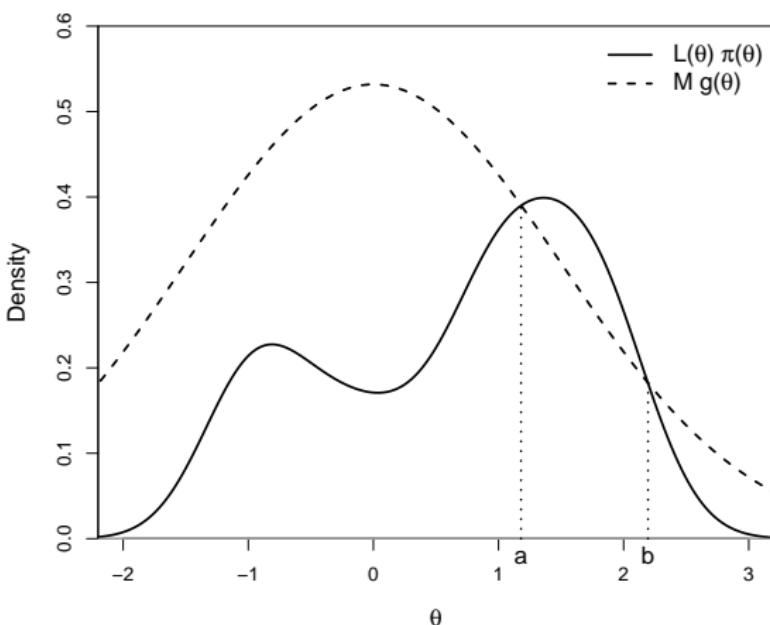


Figure 3.7: deficient rejection envelope

Then the distribution of the accepted θ 's is **not** $p(\theta | x)$, but really

$$\tilde{p}(\theta | x) = \begin{cases} \frac{L(\theta) \pi(\theta)}{\int_{S_M^c} L(\theta) \pi(\theta) d\theta + M P_g(S_M)}, & \theta \in S_M^c \\ \frac{M g(\theta)}{\int_{S_M^c} L(\theta) \pi(\theta) d\theta + M P_g(S_M)}, & \theta \in S_M \end{cases},$$

where $P_g(S_M) = \int_a^b g(\theta) d\theta$ is the probability of the set S_M . Since $\int_{S_M} L(\theta) \pi(\theta) d\theta > M P_g(S_M)$, even if $P_g(S_M)$ is small, there is no guarantee that $\tilde{p}(\theta | x)$ is close to $p(\theta | x)$. That is, only a few observed envelope violations do not necessarily imply a small inaccuracy in the posterior sample.

Example 3.9

Suppose we wish to sample values from a t_{15} distribution. (You can think of this as your posterior distribution if you like). Suppose that we cannot sample values from this distribution directly but that we can sample values from a $g = t_2$ distribution with $M = 1.5$.

R code for rejection algorithm:

```

N <- 1000
M <- 1.5    # making sure t_2 blankets the t_15 distribution
keep.thetas <- numeric(N)
samp.size <- 0
num.draws <- 0
set.seed(779)
while(samp.size < N){

  # Step 1: Generate theta value
  theta.draw <- rt(1, df = 2)

  # Step 2: Generate U ~ Uniform(0,1)
  U <- runif(1)

  # Step 3: accept or reject theta value
  if( U * M * dt(theta.draw, df = 2) < dt(theta.draw, df = 15) ){
    samp.size <- samp.size + 1
    keep.thetas[samp.size] <- theta.draw
  }
  num.draws <- num.draws + 1
}

post.mean.theta.rej <- mean(keep.thetas)

> post.mean.theta.rej
[1] -0.0547431

> N / num.draws      # acceptance rate
[1] 0.7886435

```

Here, the posterior mean of θ was estimated to be -0.055, and the acceptance rate was 0.789. The samples are given in Figure 3.8.

```
# Graph of densities and sampled theta values
plot( theta.grid, dt(theta.grid, df = 2), type = "n", cex.lab = 1.25,
      xlab = expression(theta), ylab = "", ylim = c(-0.02,0.6) )
lines( theta.grid, M * dt(theta.grid, df = 2), lty = 2 )
lines( theta.grid, dt(theta.grid, df = 15), lty = 1 )
legend( "topright", c("t(15)", "M * t(2)"), lty = c(1,2), bty = "n", cex = 1.25 )
points(keep.thetas, rep(-0.01, length(keep.thetas)), pch = 10)
points(post.mean.theta.rej, 0.01, pch = "X")
```

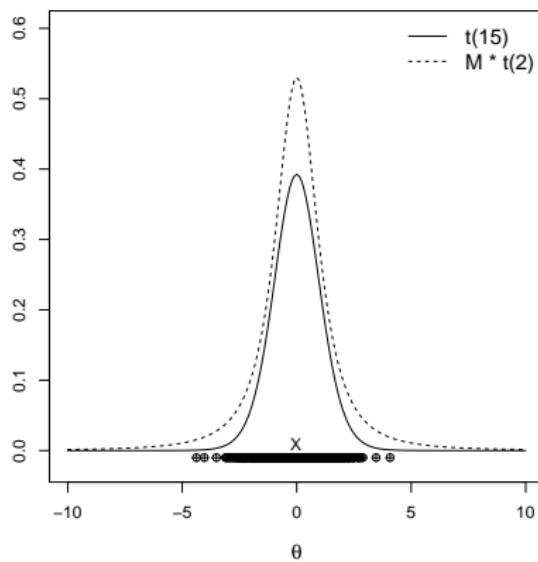


Figure 3.8:

Weighted Bootstrap

This method was presented by Smith and Gelfand (1992, *American Statistician*) and is very similar to the sampling-importance sampling resampling algorithm of Rubin (1988). Suppose an M appropriate for the rejection method is **not** readily available, but we do have a sample $\theta_1, \dots, \theta_N$ from some approximating density $g(\theta)$.

Suppose our goal is to obtain an i.i.d. sample from $c^{-1}L(\theta)\pi(\theta)$ exactly. The weighted bootstrap technique enables us to do just that. Suppose that we have used importance sampling to obtain a sample $\theta_1, \dots, \theta_N \stackrel{i.i.d.}{\sim} g(\theta)$.

Define

$$w_i = \frac{L(\theta_i) \pi(\theta_i)}{g(\theta_i)},$$

$$q_i = \frac{w_i}{\sum_{i=1}^N w_i}.$$

Now draw θ^* from the **discrete** distribution over $\{\theta_1, \dots, \theta_N\}$, which places mass q_i at θ_i .

Then θ^* is a sample from

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

with the approximation improving as $N \rightarrow \infty$. This is a **weighted bootstrap**, since instead of resampling from the set $\{\theta_1, \dots, \theta_N\}$ with equally likely probabilities of selection, we are resampling some points more than others due to the unequal weighting.

To see why this method works, note that for the standard bootstrap,

$$\begin{aligned} P(\theta^* \leq a) &= \sum_{i=1}^N \frac{1}{N} I(-\infty < \theta_i \leq a) \\ &\rightarrow \int_{-\infty}^a g(\theta) d\theta \end{aligned}$$

as $N \rightarrow \infty$, so that θ^* is approximately distributed as $g(\theta)$.

For the weighted bootstrap

$$\begin{aligned} P(\theta^* \leq a) &= \sum_{i=1}^N q_i I(-\infty < \theta_i \leq a) \\ &= \frac{\frac{1}{N} \sum_{i=1}^N w_i I(-\infty < \theta_i \leq a)}{\frac{1}{N} \sum_{i=1}^N w_i} \\ &\xrightarrow{\text{(as } N \rightarrow \infty \text{)}} \frac{E_g \left[\frac{L(\theta) \pi(\theta)}{g(\theta)} I(-\infty < \theta \leq a) \right]}{E_g \left[\frac{L(\theta) \pi(\theta)}{g(\theta)} \right]} \\ &= \frac{\int_{-\infty}^a \left[\frac{L(\theta) \pi(\theta)}{g(\theta)} g(\theta) \right] d\theta}{\int_{-\infty}^{\infty} \left[\frac{L(\theta) \pi(\theta)}{g(\theta)} g(\theta) \right] d\theta} \\ &= \frac{\int_{-\infty}^a L(\theta) \pi(\theta) d(\theta)}{\int_{-\infty}^{\infty} L(\theta) \pi(\theta) d(\theta)} = \int_{-\infty}^a p(\theta | x) d\theta , \end{aligned}$$

so that θ^* is now approximately distributed as $p(\theta | x)$, as desired.

Similar to importance and rejection sampling, we need $g(\theta) \approx p(\theta | x)$, or else a very large N will be required to obtain acceptable accuracy.

We note that if $\pi(\theta)$ is proper, it can play the role of $g(\theta)$ in any of the three methods.

Example 3.10

Suppose x_1, \dots, x_n are i.i.d. $N(\theta, \sigma^2)$ and $\tau = 1/\sigma^2$. Suppose that

$$\pi(\theta | \mu_0, \tau_0) \propto [1 + \tau_0(\theta - \mu_0)^2]^{-1},$$

Thus,

$$\begin{aligned}\theta | \mu_0, \tau_0 &\sim S_1(1, \mu_0, \tau_0^{-1}) \\ &= \text{Cauchy}(\mu_0, \tau_0^{-1}).\end{aligned}$$

Further suppose that σ^2 is known. Then the likelihood

$$L(\theta) \propto \exp \left\{ -\frac{\tau}{2} \sum (x_i - \theta)^2 \right\}$$

is maximized at $\hat{\theta} = \bar{x}$. Let $M = L(\hat{\theta})$ in the **rejection** method, and let $g(\theta) = \pi(\theta)$.

Then

$$\begin{aligned} L(\theta) \pi(\theta) \leq Mg(\theta) &\Leftrightarrow L(\theta) \pi(\theta) \leq L(\hat{\theta}) \pi(\theta) \\ &\Leftrightarrow L(\theta) \leq L(\hat{\theta}) . \end{aligned}$$

So, we simply generate $\theta_j \sim \pi(\theta)$, $U \sim \text{Uniform}(0, 1)$, and accept θ_j if

$$MUg(\theta_j) < L(\theta_j) \pi(\theta_j) ,$$

that is, if

$$\begin{aligned} U &< \frac{L(\theta_j)\pi(\theta_j)}{Mg(\theta_j)} \\ &= \frac{L(\theta_j)}{L(\hat{\theta})} . \end{aligned}$$

We note that this ratio is the probability of accepting a θ_j candidate. Hence, this approach will be quite inefficient unless $\pi(\theta)$ is **not too flat** relative to the likelihood $L(\theta)$, so that most of the candidates have a reasonable chance of being accepted. Unfortunately, this will not normally be the case, since in most cases, the data will carry much more information about θ than the prior.

In these instances, setting $g = \pi$ will also result in poor performance by importance sampling or the weighted bootstrap, since g will be a poor approximation to $L\pi$. As a result, g should be chosen as the prior π as a **last resort**, when methods of finding a better approximation have failed.

Example 3.11

We may use the weighted bootstrap procedure to sample from the t_{15} density discussed earlier. After implementing the importance sampling code, we need only do add the following steps:

R code for weighted bootstrap algorithm:

```
# Obtain theta values and weights from importance sampling
N <- 1000
set.seed(779)
thetas <- rt(N, df = 2)
w.theta <- dt(thetas, df = 15) * 1 / dt(thetas, df = 2)

# Compute normalized weights
qi <- w.theta/sum(w.theta)

# Sample m = 500 theta values with replacement
m <- 500
theta.stars <- sample(thetas, size = m, replace = TRUE, prob = qi)

# Compute posterior mean
post.mean.theta.wtboot <- mean(theta.stars)

> post.mean.theta.wtboot
[1] -0.04204751
```

The posterior mean of θ from the weighted bootstrap method is -0.042. A plot of the sample obtained is provided in Figure 3.9.

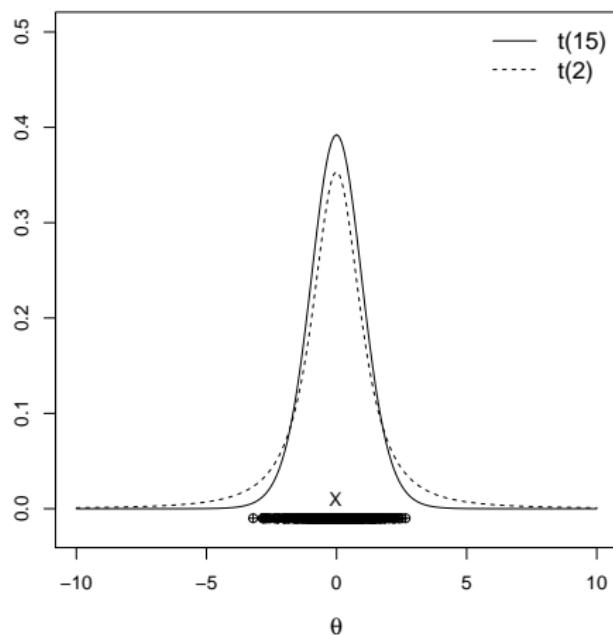


Figure 3.9:

Example 3.12: noniterative Monte Carlo methods

Consider a likelihood that is a mixture of two normals, say

$$L(\theta) = 0.4 N(\theta - 1, 0.95^2) + 0.6 N(\theta + 1, 0.95^2).$$

Suppose the prior on θ is flat such that

$$\pi(\theta) \propto 1,$$

and further suppose there is only one data point $x = 1$.

We want to sample from the posterior $p(\theta|x)$ using

- (i) importance sampling,
- (ii) rejection sampling,
- (iii) weighted bootstrap.

(i) Importance Sampling:

We will draw N values of θ from the t distribution with 2 degrees of freedom and dispersion 2.5^2 .

```
# Posterior distribution
x <- 1
dmixnorm <- function(theta, x){
  part1 <- dnorm(theta, mean = x - 1, sd = 0.95)
  part2 <- dnorm(theta, mean = x + 1, sd = 0.95)
  return(0.4 * part1 + 0.6 * part2)
}

# Draw N theta values from the t distribution with 2 df and dispersion 2.5^2
N <- 1000
sqrt.disp <- 2.5
set.seed(779)
thetas <- sqrt.disp * rt(N, df = 2)

# To compute posterior mean of theta (first moment), set f1 = theta
f1.theta <- thetas
# Compute the importance sampling weights
w.theta <- dmixnorm(thetas, x) * 1 / dt(thetas/sqrt.disp, df = 2)
# Estimate the posterior expectation
numerator <- sum(f1.theta * w.theta) / N
denominator <- sum(w.theta) / N
post.mean.theta <- numerator / denominator

> post.mean.theta
[1] 1.174445
```

The estimate of the posterior mean of θ is 1.174.

```
# To compute 2nd moment, set f2 = theta^2
f2.theta <- thetas^2
numerator <- sum(f2.theta * w.theta) / N
post.mean.theta2 <- numerator / denominator
# Estimate the posterior expectation
post.var.theta <- post.mean.theta2 - (post.mean.theta)^2

> post.var.theta
[1] 1.899755
```

The estimate of the posterior variance of θ is 1.900.

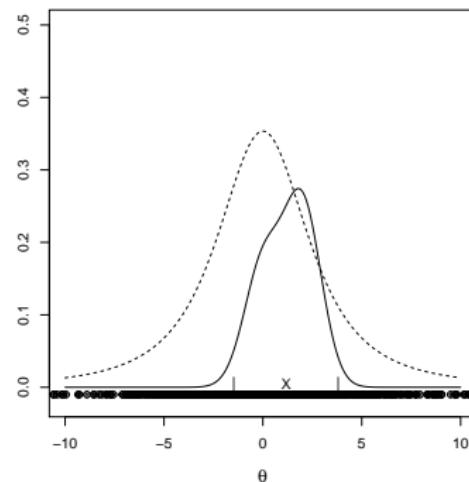


Figure 3.10:

(ii) Rejection Sampling:

We will draw N values of θ from the t distribution with 2 degrees of freedom and dispersion 2.5^2 , and we set $M = 1.25$.

```
# Draw N theta values from g, a t distribution with 2 df and dispersion 2.5^2
N <- 1000
M <- 1.25
sqrt.disp <- 2.5
keep.thetas <- numeric(N)
samp.size <- 0
num.draws <- 0
set.seed(779)
while(samp.size < N){

  # Step 1: Generate theta value
  theta.draw <- sqrt.disp * rt(1, df = 2)

  # Step 2: Generate U ~ Uniform(0,1)
  U <- runif(1)

  # Step 3: accept or reject theta value
  if( U * M * sqrt.disp * dt(theta.draw, df = 2) < dmixnorm(theta.draw, 1) ){
    samp.size <- samp.size + 1
    keep.thetas[samp.size] <- theta.draw
  }
  num.draws <- num.draws + 1

}
post.mean.theta.rej <- mean(keep.thetas)
```

```
> post.mean.theta.rej
```

```
[1] 1.716685
```

```
> N / num.draws # acceptance rate
```

```
[1] 0.3034901
```

The posterior mean of θ was estimated to be 1.717, and the acceptance rate was 0.303.

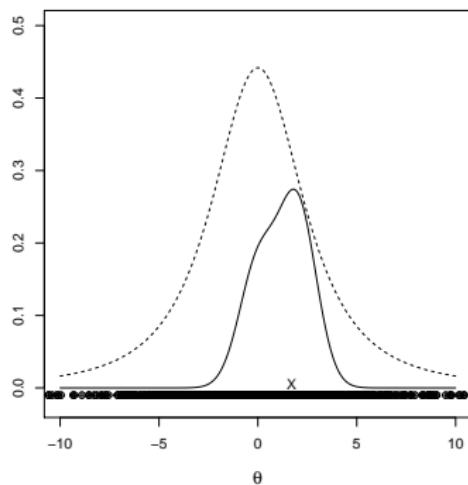


Figure 3.11:

(iii) Weighted Bootstrap:

We will draw N values of θ from the t distribution with 2 degrees of freedom and dispersion 1.5^2 .

```
# Posterior distribution
x <- 1
dmixnorm <- function(theta, x){
  part1 <- dnorm(theta, mean = x - 1, sd = 0.95)
  part2 <- dnorm(theta, mean = x + 1, sd = 0.95)
  return(0.4 * part1 + 0.6 * part2)
}

# Draw N theta values from the t distribution with 2 df and dispersion 1.5^2
N <- 1000
sqrt.disp <- 1.5
set.seed(779)
thetas <- sqrt.disp * rt(N, df = 2)

# Compute weights (same as importance sampling weights)
w.theta <- dmixnorm(thetas, x) * 1 / dt(thetas/sqrt.disp, df = 2)

# Compute bootstrap probabilities
qi <- w.theta / sum(w.theta)

# Sample with replacement from thetas using the probabilities "qi"
m <- 500
new.thetas <- sample(thetas, size = m, replace = TRUE, prob = qi)

# Compute posterior statistics
post.mean.theta <- mean(new.thetas)
post.var.theta <- var(new.thetas)
```

```
> post.mean.theta  
[1] 1.148638  
  
> post.var.theta  
[1] 1.766787
```

The posterior mean of θ was estimated to be 1.149, and the estimate of the posterior variance of θ is 1.767.

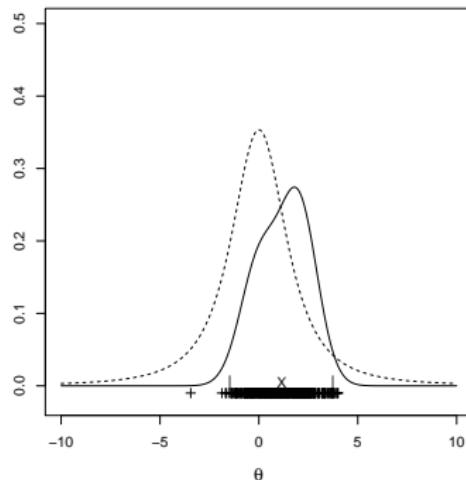


Figure 3.12:

Markov Chain Monte Carlo Methods

Importance sampling, rejection sampling, and the weighted bootstrap are all **noniterative** methods. They draw a sample of size N , and stop. Hence there is no notion of the algorithm “converging” – we simply require N sufficiently large. But for many problems, especially **high dimensional** ones, it may be quite difficult or even impossible to find an importance sampling density (or envelope function) which is an acceptably accurate approximation to the log posterior, but still easy to sample from. This facilitates a need for new algorithms.

Gibbs Sampling

Suppose we have a collection of k random variables denoted by $U = (U_1, \dots, U_k)$. We assume that the **full conditional distributions**

$$\{p(u_i | u_j, j \neq i), i = 1, \dots, k\}$$

are available for sampling. Here, “available” means that samples may be generated by some method.

Thus, we do **not** require the one dimensional conditional distributions $[U_i | U_j, j \neq i], i = 1, \dots, k$ to have a closed form, but we only need to be able to write them up to a normalizing constant.

Under mild conditions (see Besag, 1974), the one-dimensional conditional distributions **uniquely** determine the full joint distribution $[U_1, \dots, U_k]$, and hence all marginal distributions $[U_i]$, $i = 1, \dots, k$. The algorithm proceeds as follows.

Gibbs Sampling Algorithm

- 0) Suppose we have a set of **arbitrary** starting values $\{U_1^{(0)}, \dots, U_k^{(0)}\}$.
- 1) Draw $U_1^{(1)}$ from $[U_1 | U_2^{(0)}, \dots, U_k^{(0)}]$
- 2) Draw $U_2^{(1)}$ from $[U_2 | U_1^{(1)}, U_3^{(0)}, \dots, U_k^{(0)}]$
- \vdots
- k) Draw $U_k^{(1)}$ from $[U_k | U_1^{(1)}, \dots, U_{k-1}^{(1)}]$

This completes one iteration of the Gibbs sampler. Thus after one iteration, we have $(U_1^{(1)}, \dots, U_k^{(1)})$. After t such iterations, we would obtain $(U_1^{(t)}, \dots, U_k^{(t)})$.

Remark 3.2

For notational convenience, we let $[X \mid Y]$ denote the conditional distribution of X given Y . We are led to the following theorem.

Theorem 3.2

For the Gibbs sampling algorithm outlined above,

- a) $(U_1^{(t)}, \dots, U_k^{(t)}) \xrightarrow{d} [U_1, \dots, U_k]$ as $t \rightarrow \infty$.
- b) The convergence in part a) is **exponential** in t .

For proofs of these theorems, see Geman and Geman (1984, *IEEE*) and Schervish & Carlin (1992, *Journal of Computational and Graphical Statistics*). The original paper on the Gibbs sampler was by Geman and Geman (1984). Gelfand and Smith (1990, *JASA*) exposed this algorithm to statisticians in the context of Bayesian inference.

Thus, in the context of Bayesian analysis, we are interested in sampling from the joint posterior distribution $[U_1, \dots, U_k \mid x]$. The Gibbs sampler then requires draws from each of the univariate conditional distributions $p(u_i \mid u_j, x, i \neq j)$.

A marginal density estimate of U_i is given by

$$\hat{p}(u_i | x) = \frac{1}{m} \sum_{j=1}^m p(u_i | u_{1,j}^{(t)}, u_{i-1,j}^{(t)}, u_{i+1,j}^{(t)}, \dots, U_{k,j}^{(t)}, x).$$

This type of estimate results in a less variable estimate of $p(u_i | x)$ than could be obtained by kernel density estimation.

Example 3.13: 2-Dimensional Case

Suppose we want to sample from the joint distribution of (X, Y) . The Gibbs sampling algorithm proceeds as follows.

- 0) Start with an initial $(X^{(0)}, Y^{(0)})$.
- 1) Draw $X^{(1)}$ from $[X | Y^{(0)}]$.
- 2) Draw $Y^{(1)}$ from $[Y | X^{(1)}]$.

After t iterations, we have $(X^{(t)}, Y^{(t)})$.

The t iterations produce

$$(X^{(1)}, Y^{(1)}), \dots, (X^{(t)}, Y^{(t)}).$$

Example 3.14

Suppose x_1, \dots, x_n are i.i.d. $N(\mu, \sigma^2)$, $\tau = 1/\sigma^2$, $\pi(\mu, \tau) \propto \tau^{-1}$. We want to devise the Gibbs sampler to sample from the joint posterior density of (μ, τ) .

We have

$$\begin{aligned}\mu | x, \tau &\sim N\left(\bar{x}, \frac{1}{n\tau}\right) \\ \tau | x, \mu &\sim \text{gamma}\left(\frac{n}{2}, \frac{\sum(x_i - \mu)^2}{2}\right).\end{aligned}$$

To obtain joint posterior samples from $[\mu, \tau | x]$ we

- 0) Start with arbitrary $(\mu^{(0)}, \tau^{(0)})$.
- 1) Sample $\mu^{(1)}$ from $[\mu | x, \tau^{(0)}]$.
- 2) Sample $\tau^{(1)}$ from $[\tau | x, \mu^{(1)}]$.

Repeat until we have t iterations yielding

$$(\mu^{(1)}, \tau^{(1)}), \dots, (\mu^{(t)}, \tau^{(t)}).$$

Example 3.15: Logistic Regression

Suppose y_1, \dots, y_n are independent $\text{Binomial}(1, p_i)$, where

$$p_i = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}, \quad i = 1, \dots, n.$$

We have

$$L(\beta) = p(y | x, \beta) = \exp \left[\sum_{i=1}^n \{y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp\{\beta_0 + \beta_1 x_i\})\} \right].$$

Suppose we take $\pi(\beta_0, \beta_1) \propto 1$. Thus, we have

$$p(\beta_0, \beta_1 | x, y) \propto L(\beta) = \exp \left[\sum_{i=1}^n \{y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp\{\beta_0 + \beta_1 x_i\})\} \right].$$

How do we set up the Gibbs sampler for sampling from the joint posterior of $[\beta_0, \beta_1 | x, y]$?

The univariate conditionals do **not** have a closed form.

$$p(\beta_0 | \beta_1, x, y) \propto \exp \left[\sum_{i=1}^n \{y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp\{\beta_0 + \beta_1 x_i\})\} \right],$$

where β_1 is treated as **fixed**. Now

$$p(\beta_1 | \beta_0, x, y) \propto \exp \left[\sum_{i=1}^n \{y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp\{\beta_0 + \beta_1 x_i\})\} \right],$$

where β_0 is treated as fixed.

Note that the univariate conditionals for β_0 and β_1 have the **exact same functional form**, with different variables being treated as fixed. In general, the univariate conditional distribution is **always** proportional to the joint distribution. That is, if $\theta = (\theta_1, \dots, \theta_k)$, then

$$p(\theta_i | \theta_j, x, j \neq i) \propto L(\theta) \pi(\theta), \quad i = 1, \dots, k.$$

For the logistic regression problem, since the univariate conditionals **do not** have a closed form, we need a rejection algorithm, importance algorithm, or weighted bootstrap to sample from these univariate conditionals. Rejection algorithms seem to be the most popular and have been nicely developed by Gilks and Wild (1992, *Applied Statistics*) and Gilks *et al.* (1994, *Applied Statistics*).

Thus in many situations, we need to use a rejection algorithm **within** a Gibbs sampling algorithm. For details and several examples on Gibbs sampling, see Gelfand and Smith (1990), Gilks *et al.* (1993, *JRSS-B*), Gelfand *et al.* (1990, *JASA*), Wakefield *et al.* (1994, *Applied Statistics*).

Introduction to Stan

Stan is a probabilistic programming language for Bayesian statistical inference written in C++ that interfaces with other data analysis languages (R, Python, shell, MATLAB, Julia, Stata).

In R, Stan can be called using either the `rstan` or `rstanarm` packages. The `rstan` package allows for greater flexibility when specifying models, and the user specifies the model details in a `.stan` file.

The `rstanarm` package estimates previously compiled regression models using the `rstan` package.

The *.stan* file should contain the following sections of code:

- ▶ data: list the names and types of all variables along with any constraints,
- ▶ parameters: list the names and types of all parameters along with any constraints,
- ▶ model: list the likelihood of the data and all prior distributions.

An optional section of code titled “transformed parameters” can be included to list any transformations of parameters (e.g., $\tau = 1/\sigma^2$).

The model can be written out in standard regression notation using either scalar or matrix form.

See the Stan documentation on the [Stan website](#) for additional help.

Example 3.16: Infant Systolic Blood Pressure

Consider the following data for infant systolic blood pressure (mm Hg), age (in days), and birthweight (in oz).

Birthweight	Age	SBP
135	3	89
120	4	90
100	3	83
105	2	77
130	4	92
125	5	98
125	2	82
105	3	85
120	5	96
90	4	95
120	2	80
95	3	79
120	3	86
150	4	97
160	3	92
125	3	88

We wish to consider the model $SBP_i = \beta_0 + \beta_1 BWT_i + \beta_2 AGE_i + \epsilon_i$. Here we let $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 is the variance and $\tau = \frac{1}{\sigma^2}$.

We will use Stan to set up a Gibbs sampler to sample from the joint posterior distribution of $[\boldsymbol{\beta}, \sigma^2 | y, X]$.

Stan Code for Example 3.16

We will use noninformative priors for τ and the regression coefficients, specifically, $\tau \sim \text{gamma}(.001, .001)$, $\beta_0 \sim N(0, s_0 = 100)$, $\beta_1 \sim N(0, s_0 = 100)$, $\beta_2 \sim N(0, s_0 = 100)$, where s_0 is the standard deviation hyperparameter specified in the normal priors for the regression parameters. Note that $\sigma^2 = \frac{1}{\tau} \sim \text{IG}(.001, .001)$ (Inverse Gamma).

The data are contained in the file *sbp_example.dat*, which is reproduced here:

```
135 3 89
120 4 90
100 3 83
105 2 77
130 4 92
125 5 98
125 2 82
105 3 85
120 5 96
90 4 95
120 2 80
95 3 79
120 3 86
150 4 97
160 3 92
125 3 88
```

Stan code contained in the file *sbp-example.stan*:

```
data {  
    // Define variables in data  
    int N;           // Number of observations  
    vector[N] sbp;   // systolic blood pressure (continuous outcome)  
    vector[N] bwt;   // birthweight (covariate)  
    vector[N] age;   // age (covariate)  
}  
  
parameters {  
    // Define parameters  
    real beta0;  
    real beta1;  
    real beta2;  
    real<lower=0> sigma2;  
}  
  
transformed parameters {  
    // Mean  
    vector[N] mu;  
    mu = beta0 + beta1 * bwt + beta2 * age;  
}  
  
model {  
    // Prior distributions  
    // Specify standard deviation (not variance or precision) in normal distribution  
    beta0 ~ normal(0, 100);  
    beta1 ~ normal(0, 100);  
    beta2 ~ normal(0, 100);  
    // If tau ~ gamma(a, b) with rate parameterization, then sigma2 = 1/tau ~ inv_gamma(a, b)  
    sigma2 ~ inv_gamma(.001, .001);  
  
    // Likelihood part of Bayesian inference  
    sbp ~ normal(mu, sqrt(sigma2));  
}
```

Finally, the R code is presented here:

```
# Arrange data for Stan
# Variable names must match names in data block of "sbp_example.stan" file
sbp.data <- read.table("sbp_example.dat", header = FALSE)
N <- nrow(sbp.data)
bwt <- sbp.data[,1]
age <- sbp.data[,2]
sbp <- sbp.data[,3]

# Run Stan code to obtain 10,000 posterior draws
# (4 chains, 3000 draws each, 500 burn-in per chain)
library(rstan)
stan.mod <- stan( file = "sbp_example.stan", data = c("N", "sbp", "bwt", "age"),
                   chains = 4, iter = 3000, warmup = 500, seed = 779 )

# Save posterior draws and compute posterior summary statistics
post.samples <- as.matrix(stan.mod)

> print( stan.mod, pars = c("beta0", "beta1", "beta2", "sigma2"), digits_summary = 3 )
Inference for Stan model: sbp_example.
4 chains, each with iter=3000; warmup=500; thin=1;
post-warmup draws per chain=2500, total post-warmup draws=10000.

      mean se_mean    sd   2.5%   25%   50%   75% 97.5% n_eff Rhat
beta0  53.284  0.085 4.953 42.926 50.278 53.351 56.398 63.102 3413 1.002
beta1   0.127  0.001 0.037  0.054  0.102  0.126  0.150  0.204 3683 1.001
beta2   5.900  0.010 0.744  4.420  5.426  5.898  6.366  7.383 5208 1.000
sigma2  7.372  0.058 3.508  3.233  5.055  6.569  8.725 16.262 3648 1.001
```

Our estimate for the posterior mean of the birthweight coefficient is 0.127 with a 95% credible set given by (0.054, 0.204). The estimate for the posterior mean of the age coefficient is 5.900, with a 95% credible set given by (4.420, 7.383).

A frequentist analysis was carried out in SAS. Does this agree with the results obtained from the Bayesian analysis?

Table 3.1: Parameter estimates from frequentist SAS analysis

Variable	DF	Parameter Estimate	Standard Error	T for H_0 Parameter = 0	Prob > T
INTERCEPT	1	53.450194	4.53188859	11.794	0.0001
BWT	1	0.125583	0.03433620	3.657	0.0029
AGE	1	5.887719	0.68020515	8.656	0.0001

Metropolis-Hastings Algorithm

Like the Gibbs sampling algorithm, the Metropolis-Hastings algorithm is an MCMC (Markov chain Monte Carlo) method. The originator of the algorithm is Metropolis *et al.* (1953, *Chemical Physics*). Hastings (1970, *Biometrika*) introduced this algorithm for statistical problems.

Suppose we wish to sample from the joint distribution $[U_1, \dots, U_k] = [U]$, where $U = (U_1, \dots, U_k)$. Denote the density of U by $p(u)$.

Let $q(v, u)$ be a density in u such that $q(u, v) = q(v, u)$. The function q is called a **candidate** or **proposal** density. We generate values as follows.

Metropolis Algorithm

1. Draw $v \sim q(\cdot, u)$, where $u = U^{(t-1)}$, the current state of the Metropolis algorithm.
2. Compute the odds ratio

$$r = \frac{p(v)}{p(u)} = \frac{L(v)\pi(v)}{L(u)\pi(u)}$$

3. If $r \geq 1$, set $U^{(t)} = v$ (acceptance)
4. If $r < 1$, set

$$U^{(t)} = \begin{cases} v & \text{with probability } r \\ u & \text{with probability } 1 - r \end{cases}$$

(Rejection)

Theorem 3.3

For the Metropolis algorithm, under mild conditions, $[U^{(t)}] \rightarrow [U]$ as $t \rightarrow \infty$.

Remark 3.3

The Gibbs sampling algorithm and the Metropolis-Hastings algorithm are **Markov chain Monte Carlo** algorithms since the samples that are generated by these algorithms are samples from a certain **Markov chain**.

For continuous parameter settings, the most convenient choice for q is a $N(\theta^{(t-1)}, \tilde{\Sigma})$, where $\tilde{\Sigma}$ can be taken as

$$\tilde{\Sigma} = - \left[\frac{\partial^2 \log[p^*(\theta|x)]}{\partial \theta \partial \theta'} \right]^{-1} \Big|_{\theta=\theta^{(t-1)}},$$

where $p^*(\theta|x) = L(\theta)\pi(\theta)$. We note that this choice of q is easily sampled and clearly symmetric in v and $\theta^{(t-1)}$.

A simple but important generalization of the Metropolis algorithm was provided by Hastings (1970, *Biometrika*).

Hastings (1970) **drops** the requirement that $q(u, v)$ be symmetric, and redefined the odds ratio as

$$r = \frac{p(v) q(u, v)}{p(u) q(v, u)}.$$

With this modification, it can be shown that this algorithm converges to the required “target” distribution for any candidate density q . Chib and Greenberg (1995, *American Statistician*) give an excellent discussion of the Metropolis-Hastings algorithms and discuss choices of $q(u, v)$.

Example 3.17

Consider the following data from Bliss (1935, *Annals of Applied Biology*). The data record the number of adult flour beetles killed after five hours of exposure to various levels of gaseous carbon disulphide (CS_2).

Dosage w_i	# killed y_i	# exposed n_i
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

Consider the model

$$P(\text{death} \mid w_i) \equiv h(w_i) = \left[\frac{\exp(x_i)}{1 + \exp(x_i)} \right]^{m_1}$$

where $m_1 > 0$, w_i is the covariate (dose), and $x_i = \frac{w_i - \mu}{\sigma}$, where $\mu \in R^1$ and $\sigma^2 > 0$.

For the prior distributions, we take

$$\begin{aligned} m_1 &\sim \text{gamma}(a_0, b_0^{-1}) , \\ \mu &\sim N(c_0, d_0) , \\ \sigma^2 &\sim IG(e_0, f_0^{-1}) \quad (\text{Inverse Gamma}) \end{aligned}$$

Note: If $z \sim IG(\frac{a}{2}, \frac{b}{2})$, then

$$p(z \mid a, b) = \frac{(b/2)^{a/2}}{\Gamma(a/2)} z^{-\left(\frac{a}{2} + 1\right)} \exp\left\{-\frac{b}{2z}\right\} ,$$

and thus $v = \frac{1}{z} \sim \text{gamma}(\frac{a}{2}, \frac{b}{2})$.

Thus if $\tau = 1/\sigma^2$ and if $\sigma^2 \sim IG(e_0, f_0)$, then $\tau \sim \text{gamma}(e_0, f_0)$.

We take m_1, μ, σ^2 to be independent a priori. The joint posterior is given by

$$\begin{aligned} p(\mu, \sigma^2, m_1 | y) &\propto p(y | \mu, \sigma^2, m_1) \pi(\mu, \sigma^2, m_1) \\ &\propto \left\{ \prod_{i=1}^k [h(w_i)]^{y_i} [1 - h(w_i)]^{n_i - y_i} \right\} \\ &\quad \times \frac{m_1^{a_0-1}}{(\sigma^2)^{e_0+1}} \exp \left\{ -\frac{1}{2} \left(\frac{\mu - c_0}{d_0} \right)^2 - \frac{m_1}{b_0} - \frac{1}{f_0 \sigma^2} \right\}. \end{aligned}$$

Let $\theta = (\theta_1, \theta_2, \theta_3) = (\mu, \frac{1}{2} \log(\sigma^2), \log(m_1))$. This transforms the parameter space to R^3 . This will be nice if we want to work with **Gaussian proposal densities**.

Upon making this transformation, we get

$$\begin{aligned}
 p(\theta | y) &\propto \left\{ \prod_{i=1}^k [h(w_i)]^{y_i} [1 - h(w_i)]^{n_i - y_i} \right\} \\
 &\times \exp\{(a_0 - 1)\theta_3 - 2(e_0 + 1)\theta_2\} \\
 &\times \exp\left\{ -\frac{1}{2} \left(\frac{\theta_1 - c_0}{d_0} \right)^2 - \frac{\exp\{\theta_3\}}{b_0} - \frac{\exp\{-2\theta_2\}}{f_0} \right\} \\
 &\times \exp\{2\theta_2 + \theta_3\}.
 \end{aligned}$$

Numerical stability is improved by working on the log-scale for computing the Metropolis odds ratio as

$$r = \exp \left[\log p^*(v | y) - \log p^*(\theta^{(t-1)} | y) \right].$$

The hyperparameters are chosen to be

$$a_0 = .25 \text{ and } b_0 = 4$$

so that m_1 has prior mean equal to 1 (corresponding to the standard logit model) and prior standard deviation 2.

Vague priors are specified for μ and σ^2 by setting $c_0 = 2$, $d_0 = 10$, $e_0 = 2$, and $f_0 = 1000$. The latter two choices imply a prior mean of .001 and a prior standard deviation of .5 for σ^2 . We use a $N_3(\theta^{(t-1)}, \tilde{\Sigma})$ **proposal density**, with

$$\tilde{\Sigma} = \text{Diag}(.00012, .033, .10) .$$

Figure 3.13 shows the output for a single Metropolis sampling chain with 20,000 iterations for each parameter. The histograms in the figure include 10,000 samples after a **burn-in period** of 10,000 iterations.

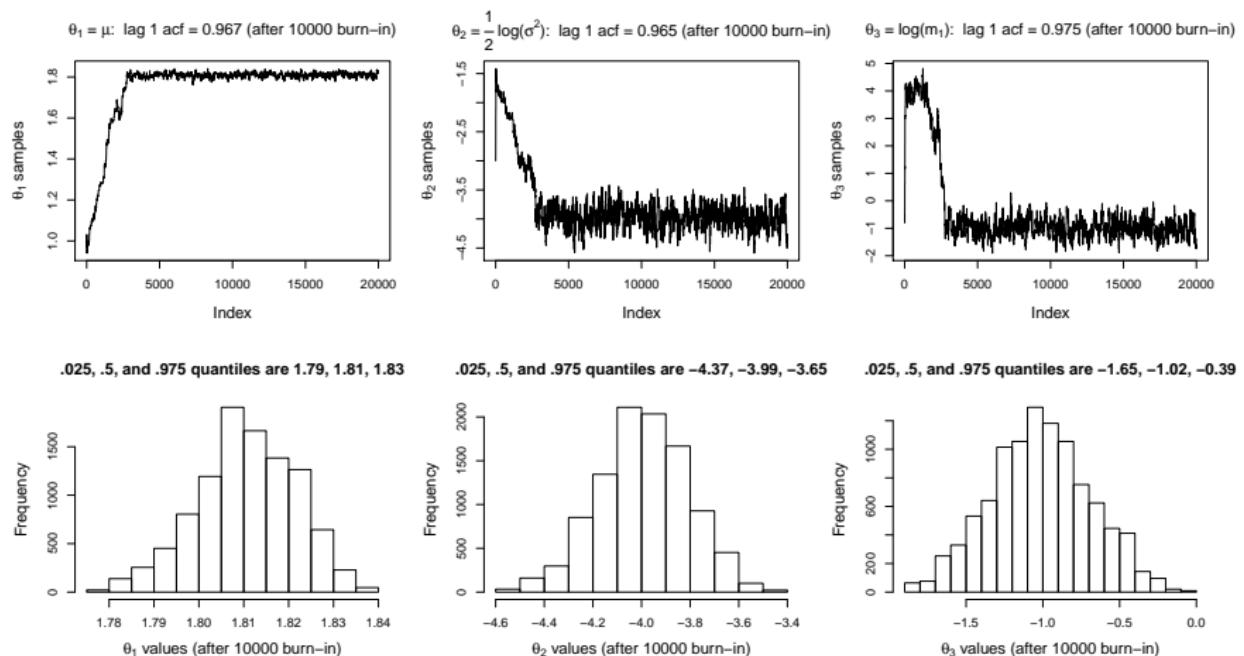


Figure 3.13:

R code for Example 3.17:

```

# Data input
w <- c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.8839)
y <- c(6,13,18,28,52,53,61,60)
n <- c(59,60,62,56,63,59,62,60)

# Log of posterior distribution
logpost <- function( w, y, n, theta1, theta2, theta3 ){
  a0 <- .25
  b0 <- 4
  c0 <- 2
  d0 <- 10
  e0 <- 2
  f0 <- 1000
  x <- (w - theta1) / exp(theta2)
  hw <- ( exp(x) / (1 + exp(x)) )^exp(theta3)
  loglikei <- y * log(hw) + (n - y) * log(1 - hw)
  likepiece <- sum(loglikei)
  part2 <- (a0 - 1) * theta3 - 2 * (e0 + 1) * theta2
  part3 <- (-1/2) * ((theta1 - c0) / d0)^2 - exp(theta3) / b0 - exp(-2 * theta2) / f0
  part4 <- 2 * theta2 + theta3
  loglikeprior <- likepiece + part2 + part3 + part4
  return(loglikeprior)
}

# q(v,u) from which we will draw samples
rmultinorm <- function(mu, Sigma){
  n2 <- 1
  p <- length(mu)
  eS <- eigen(Sigma, sym = TRUE)
  ev <- eS$values
  X <- mu + eS$vectors %*% diag( sqrt(pmax(ev,0)) ) %*% matrix(rnorm(p * n2), p)
  return(t(X))
}

```

```
# Set up Metropolis algorithm
Sigma <- diag( c(.00012, .033, .10) )
N <- 20000
theta.sample <- matrix(0, nrow = N, ncol = 3)
theta.sample[1,] <- c(1, -3, -.5)      # set starting values
num.acc <- 0    # count the number of acceptances
set.seed(779)
for(i in 2:N){
  v <- rmultinorm(theta.sample[i-1,], Sigma)
  logdmult <- logpost( w, y, n, theta.sample[i-1,1], theta.sample[i-1,2],
                        theta.sample[i-1,3] )
  dpost <- logpost( w, y, n, v[1], v[2], v[3] )
  r <- exp(dpost - logdmult)
  if(r >= 1){
    theta.sample[i,] <- c(v[1], v[2], v[3])
    num.acc <- num.acc + 1
  }
  else{
    flip <- rbinom(1, 1, r)
    if(flip == 1){
      theta.sample[i,] <- c(v[1], v[2], v[3])
      num.acc <- num.acc + 1
    }
    else{
      theta.sample[i,] <- theta.sample[i-1,]
    }
  }
}
# Calculate acceptance probability
accprob <- num.acc / N
```

```
> accprob
[1] 0.13125

# Summary of theta1, theta2, and theta3 after 5000 burn-in
burn.in <- 10000

> quantile( theta.sample[-c(1:burn.in),1], c(.025, .5, .975) )
  2.5%      50%     97.5%
1.787692 1.810737 1.830020

> quantile( theta.sample[-c(1:burn.in),2], c(.025, .5, .975) )
  2.5%      50%     97.5%
-4.365132 -3.991959 -3.646253

> quantile( theta.sample[-c(1:burn.in),3], c(.025, .5, .975) )
  2.5%      50%     97.5%
-1.6475152 -1.0223029 -0.3852342

# Lag 1 ACF values after 5000 burn-in
> acf(theta.sample[-c(1:burn.in),1])$acf[2]
[1] 0.9668653

> acf(theta.sample[-c(1:burn.in),2])$acf[2]
[1] 0.9649279

> acf(theta.sample[-c(1:burn.in),3])$acf[2]
[1] 0.9745972
```

The chains mix very slowly, as can be seen from the extremely high **lag 1 sample autocorrelations**. The reason for this slow convergence is the high correlations amongst the three parameters, estimated as $\widehat{\text{corr}}(\theta_1, \theta_2) = -.73$, $\widehat{\text{corr}}(\theta_1, \theta_3) = -.92$, and $\widehat{\text{corr}}(\theta_2, \theta_3) = .87$.

As a result, the proposal acceptance rate is low (13.1%), and convergence is slow. Convergence can be accelerated by using a nondiagonal proposal covariance matrix designed to better mimic the posterior surface. We can obtain an estimate of the posterior covariance matrix in the usual way as

$$\hat{\Sigma} = \frac{1}{N} \sum_{j=1}^N (\theta_j - \bar{\theta})(\theta_j - \bar{\theta})'$$

where j indexes the Monte Carlo samples. Now we consider

$$\tilde{\Sigma} = 2 \hat{\Sigma}$$

The results indicate improved convergence, with lower observed autocorrelations (lag 1 ACF values of 0.86, 0.85, 0.85) and a higher Metropolis acceptance rate (29.1%).

Convergence Monitoring and Acceleration

When we say that an MCMC algorithm **has converged** at iteration T , we mean that its output can be safely thought of as coming from the true **stationary distribution** of the Markov chain for all $t > T$.

The most common source of MCMC convergence difficulty is due to model **overparameterization**. When a model becomes overparameterized, the parameters become **nonidentifiable**. As an example, consider the problem of finding the posterior distribution of θ_1 , where the likelihood is defined by

$$x_i \mid \theta_1, \theta_2 \stackrel{i.i.d.}{\sim} N(\theta_1 + \theta_2, 1),$$

$i = 1, \dots, n$, and we take $\pi(\theta_1, \theta_2) \propto 1$. Here, only the sum of the two parameters is identified by the data, so **without proper** priors for θ_1 and θ_2 , their marginal posterior distributions will be improper as well. Unfortunately, a naive application of the Gibbs sampler in this setting would not reveal this problem.

In addition, models that are overparameterized typically lead to **high posterior correlations** amongst the parameters. These correlations are referred to as **cross-correlations**.

High cross-correlations will dramatically effect the movement of the Gibbs sampler through the parameter space. Even if the parameters are identifiable, high cross-correlations and high auto-correlations can lead to **slow convergence**.

One of the traditional remedies suggested for reducing high correlations amongst parameter is to **reparameterize**. Gelfand *et al.* (1995) discuss a class of transformations called **hierarchical centering**, which is quite useful for random effects models. In general, such reparameterizations are hard to find.

As an alternative to a remedy, we can develop **diagnostic statistics** that help assess whether the MCMC algorithm has converged. There has been much effort in this area. MCMC convergence diagnostics have a wide variety of characteristics.

1. **Theoretical basis** – Convergence diagnostics have been derived using a broad array of mathematical machinery and sophistication.
2. **Diagnostic Goal** – Most diagnostics address the issue of **bias**, or the distance of the estimated quantities of interest at a particular iteration from their true values under the target distribution. But a few also consider **variance**, or the precision of those estimates, which is perhaps more important from a practical standpoint.
3. **Output format** – Some diagnostics are quantitative, producing a single number summary, while others are qualitative, summarized by a graph or other display.
4. **Replication** – Some diagnostics may be implemented using only a single MCMC chain, while others require a small number of **parallel chains**.

5. **Dimensionality** – Some diagnostics consider only univariate posterior quantities, while others attempt to diagnose convergence of the full joint posterior distribution.
6. **Application** – Some diagnostics apply only to output from Gibbs samplers, some to output from any MCMC scheme, and some to a subset of algorithms somewhere between the two.
7. **Ease of use** – Generic computer code is freely available to implement some convergence diagnostics.

Convergence Diagnostic Statistics

1. Gelman & Rubin
2. Geyer
3. Geweke
4. Raftery & Lewis
5. Heidelberger and Welch
6. Autocorrelations
7. Cross-Correlations

The programming language Stan does Gibbs sampling for various types of models, and convergence diagnosis and output analysis can be performed using CODA. The R package `coda` computes several convergence diagnostic statistics using the Gibbs output from Stan.

To use Stan and CODA together in R, first save the posterior chains for each parameter by converting the `stanfit` object to a matrix (i.e., `as.matrix()`), and then convert the matrix to a MCMC object (i.e., `as.mcmc()`). You can select the parameters of interest by specifying which columns to use in either the matrix or MCMC object.

Gelman & Rubin

Perhaps the single most popular approach is due to Gelman and Rubin (1992, *Statistical Science*). They propose a convergence test based on 2 or more parallel chains, each starting from different initial values which are over-dispersed with respect to the true posterior distribution (see Gelman and Rubin (1992) for details concerning construction of over-dispersed starting distributions). Their method is based on a comparison of the within and between chain variances for each variable (essentially a classical analysis of variance).

In this approach, we

1. Run a small number (m) of parallel chains with different starting points.
2. These chains must be initially **overdispersed** with respect to the true posterior.
3. Running the m chains for $2N$ iterations each, we then attempt to check whether the **variation within the chains** for a given parameter of interest, say λ , approximately equals the **total variation across chains** during the latter N iterations.

Specifically, we monitor convergence by the estimated **scale reduction factor**,

$$\sqrt{\hat{R}} = \sqrt{\frac{N-1}{N} + \frac{m+1}{mN} \left(\frac{B}{W} \right) \left(\frac{df}{df-2} \right)},$$

where B/N is the **variance between** the means from the m parallel chains, W is the average of the m **within-chain variances**, and df is the degrees of freedom of an approximating t density to the posterior distribution.

Best results are obtained for parameters whose marginal posterior densities are approximately normal. Hence CODA transforms any variables specified to be positive or restricted to the range $(0,1)$ to the logarithmic or logit scales respectively before calculating this diagnostic.

Gelman and Rubin (1992) show that $\sqrt{\hat{R}} \rightarrow 1$ as $N \rightarrow \infty$. Thus values of $\sqrt{\hat{R}}$ close to 1 suggest good convergence.

The Gelman and Rubin diagnostics reported by CODA are the 50% and 97.5% quantiles of the sampling distributions for this scale reduction or shrink factor. Note that the quantiles are estimated from the second half of each chain only. If both quantiles are approximately 1.0, effective convergence may be diagnosed (i.e., samples from the second half of each chain may be assumed to have arisen from the stationary distribution). In this case, summary statistics, density estimates, etc., may be calculated by combining the latter 50% of iterates from all chains.

This approach has been criticized because

1. It's a univariate diagnostic. It must be applied to each parameter.
Thus, if $\theta = (\theta_1, \dots, \theta_p)$, we need to compute $\sqrt{\hat{R}_{\theta_j}}$, $j = 1, \dots, p$.
2. The approach focuses solely on the **bias component** of convergence, providing no information as to the accuracy of the resulting posterior estimates.
3. The method relies heavily on the user's ability to find a starting distribution that is actually **overdispersed** with respect to the true posterior distribution of λ , a condition we can't really check without knowledge of the latter.

To implement Gelman and Rubin (1992)'s convergence diagnostic in R, save the MCMC chains into a list and convert to a `mcmc.list` object (i.e., `as.mcmc.list()`) and use the `gelman.diag()` function in the `coda` package.

Geyer

Geyer (1992, *Statistical Science*) recommends a dramatically different approach, wherein we run a **single** chain, and focuses not on the bias but the **variance** of the resulting estimates. For example, letting

$\hat{\lambda}_N = \frac{1}{N} \sum_{t=1}^N \lambda^{(t)}$, we have, by the central limit theorem,

$$\sqrt{N} (\hat{\lambda}_N - \lambda) \xrightarrow{d} N(0, \sigma^2) ,$$

where σ^2 can be estimated by

$$\hat{\sigma}_N^2 = \sum_{t=0}^{\infty} w_N(t) \hat{\gamma}_{N,t} .$$

Here, $w_N(t)$ is a weight function called a **lag window**, and $\hat{\gamma}_{N,t}$ is an estimate of the lag t autocovariance function, namely

$$\hat{\gamma}_{N,t} = \frac{1}{N} \sum_{i=1}^{N-t} (\lambda^{(i)} - \hat{\lambda}_N) (\lambda^{(i+t)} - \hat{\lambda}_N) .$$

The lag window satisfies $0 \leq w_N(t) \leq 1$ and is used to downweight the large-lag terms. An approximate 95% confidence interval for $E(\lambda | x)$ is thus obtainable as

$$\hat{\lambda}_N \pm 1.96 \frac{\hat{\sigma}_N}{\sqrt{N}} .$$

The previous formulae assume **no discarding** of early samples (i.e., no burn-in). Geyer argues that less than 1% of the run will normally be a sufficient burn-in period whenever the run is long enough to give much precision. Recall that Gelman and Rubin discard the first 50% of each chain.

Geyer's approach relies on more sophisticated mathematics, addresses the variance goal, and avoids parallel chains or any notion of "initial overdispersion." However, it **doesn't** generalize well beyond posterior moments, that is, to posterior quantiles or density estimates for λ , and it does not concern itself with actually looking for bias.

A computationally simpler method of estimating $\text{Var}(\hat{\lambda}_N)$ is through **batching**. Suppose a single long run of length $N = mk$ is conducted. Divide the run into m **successive batches** of length k with batch means B_1, \dots, B_m . We have

$$\hat{\lambda}_N = \bar{B} = \frac{1}{m} \sum_{i=1}^m B_i .$$

We then have the variance estimate

$$\hat{V} \equiv \widehat{\text{Var}}(\hat{\lambda}_N) = \frac{1}{m(m-1)} \sum_{i=1}^m (B_i - \hat{\lambda}_N)^2 ,$$

provided that k is large enough so that the correlation between batches is negligible, and m is large enough to reliably estimate $\text{Var}(B_i)$. The confidence interval for $E(\lambda | x)$ is then given by

$$\hat{\lambda}_N \pm t_{(m-1,.025)} \sqrt{\hat{V}} ,$$

where $t_{(m-1,.025)}$ is the upper .025 point of a t distribution with $m-1$ degrees of freedom.

A final solution to the problem of estimating $\text{Var}(\hat{\lambda}_N)$ would be to simply use the sample variance of the k^{th} iteration output from m independent parallel chains. However, this approach is horribly wasteful, discarding $\frac{k-1}{k}$ of the samples.

Geweke

Geweke (1992) proposes a convergence diagnostic based on standard time-series methods. The test is appropriate for use with single chains when convergence of the mean of some function of the sampled variables is of interest. For each (function of the) variable, the chain is divided into 2 “windows” containing the first $x\%$ (CODA default is 10%) and last $y\%$ (CODA default is 50%) of the iterates. If the whole chain is stationary, the means of the values early and late in the sequence should be similar. Geweke’s approach involves calculation of the sample mean and asymptotic variance in each window, the latter being determined by spectral density estimation.

His convergence diagnostic Z is the difference between these 2 means divided by the asymptotic standard error of their difference. As the chain length $\rightarrow \infty$, the sampling distribution of $Z \rightarrow N(0, 1)$ if the chain has converged. Hence values of Z which fall in the extreme tails of a standard normal distribution suggest that the chain was not fully converged early on (i.e., during the 1st window).

To implement Geweke (1992)’s convergence diagnostic in R, use the `geweke.diag()` function in the `coda` package.

Raftery & Lewis

Raftery and Lewis (1992b)'s method applies to single chains. It is intended both to detect convergence to the stationary distribution and to provide bounds for the accuracy of the estimated quantiles of functions of variables of interest. The user must specify the quantile to be estimated (default is 2.5 percentile), the desired degree of accuracy (the CODA default is ± 0.005), and the required probability of attaining this degree of accuracy (the CODA default is 0.95).

The CODA output reports N_{min} , the minimum number of iterations that would be needed to estimate the specified quantile to the desired precision if the samples in the chain were independent. This is a theoretical value based on the binomial variance and provides a lower bound for the run-length of the Gibbs sampler. (Note however that in the unusual event that consecutive samples in the output chain are *negatively* correlated, fewer than N_{min} iterations will be needed). N_{min} will increase as the required probability and degree of accuracy increase.

Somewhat counter-intuitively, it also will be larger when estimating quantiles close to the median compared to more extreme quantiles. If the chain length of the Stan output is less than N_{min} for the quantile, accuracy and probability values currently specified in CODA, the program will not compute the Raftery & Lewis diagnostics but will issue an error message stating the value of N_{min} . If sufficient Stan iterations are available CODA will report

1. N , the total number of iterations that should be run for each variable,
2. M , the number of initial iterations to discard as the “burn-in,”
3. $I = \frac{M+N}{N_{min}}$, which measures the extent to which autocorrelation inflates the required sample size.

Raftery and Lewis (1992a) suggest that $I > 5.0$ often indicates problems. In this case, reparametrizations of the model is advised.

To implement Raftery and Lewis (1992b)'s convergence diagnostic in R, use the `raftery.diag()` function in the `coda` package.

Heidelberger and Welch

Heidelberger and Welch (1983) devised a method for detecting an initial transient in simulated sequences of discrete events, but which is also appropriate for use as a convergence diagnostic for the output of Gibbs samplers. Their idea is based on Brownian bridge theory and uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values for each variable form a stationary process. If the null hypothesis is rejected for a given variable, a further 10% of iterations are discarded. This process is repeated until either a portion of the chain (of length $\geq 50\%$ of the total number of iterations) passes the stationary test, or 50% of the iterations have been discarded and the null hypothesis is still rejected. If the latter occurs, CODA reports the Cramer-von-Mises statistic and indicates that the stationarity test was failed. Missing values (NA) are shown in all other columns of the output table for that variable. This indicates that a longer Stan run is needed in order to achieve converge.

If the stationarity test passes, CODA reports the number of iterations to keep (which are diagnosed to arise from a stationary process), the number of initial iterations to discard, and the Cramer-von-Mises statistic.

A halfwidth test is then carried out as follows: for each variable, the portion of the chain which passed the stationarity test is used to estimate the asymptotic standard error of the mean via a time-series method. CODA reports the sample mean of the retained iterates and the halfwidth of the associated 95% confidence interval for this mean (i.e. $1.96 \times$ asymptotic standard error). If the halfwidth is less than ϵ times the sample mean (where ϵ is a small fraction), the halfwidth test is passed and the retained sample is deemed to estimate the posterior mean with accepting precision. If the halfwidth test is failed, this implies that a longer Stan run is needed to increase the accuracy of the posterior estimates for a given variable. The CODA default for ϵ is 0.1.

To implement Heidelberger and Welch (1983)'s convergence diagnostic in R, use the `heidel.diag()` function in the `coda` package.

Autocorrelations

The `autocorr.diag()` function in the R package `coda` produces a table of autocorrelations within each chain for each monitored variable at lags of 1, 5, 10, and 50. High autocorrelations within chains indicate slow mixing and, usually, slow convergence. This will be characterized by plots of sample traces which “snake” slowly up and down, as opposed to showing more rapid fluctuations over the sample space.

Reparametrizations may help to reduce autocorrelations. Alternatively, it may be necessary to increase the thinning interval to say, every 5th or 10th iteration, before calculating summary statistics and density estimates, in order to achieve a less highly correlated sample.

Cross-Correlations

Correlations between monitored variables in each chain. High correlations among parameters are associated with slow convergence and may indicate a need for reparametrization of the model. Use the `crosscorr()` function in the R package `coda`.

Cowles and Carlin (1996, *JASA*) give a nice overview of several convergence diagnostic statistics.

In summary, here is a potential convergence diagnostic strategy.

1. Run a few (3 to 5) parallel chains, with starting points drawn from a distribution believed to be overdispersed with respect to the stationary distribution (say, covering ± 3 prior standard deviations from the prior mean).
2. Visually inspect these chains by overlaying their sampled values on a common graph for each parameter, or, for very high dimensional models, a representative subset of the parameters.
3. Annotate each graph with the Gelman & Rubin (1992) statistics and lag 1 autocorrelations. Large Gelman and Rubin (1992) statistics may arise from either slow mixing or multimodality.
4. Investigate cross correlations among parameters suspected of being nearly nonidentifiable.

An area closely related to convergence diagnosis is that of convergence acceleration. Reparameterizations can often improve a model's correlation structure and hence speed up convergence. The closer the correlations are to 0, the faster the convergence. See Gelfand *et al.* (1995, *Bayesian Statistics 5*) for techniques on reparameterization.

Example 3.18: Infant Systolic Blood Pressure (cont.)

Compute convergence diagnostic statistics on the posterior draws for β_0 , β_1 , β_2 , and σ^2 using the results from the Gibbs sampler in Example 3.16.

```
# Extract posterior samples beta0, beta1, beta2, sigma2
# "stan.mod" is the stanfit object from Example 3.16
library(coda)
post.samples <- as.matrix(stan.mod)[,1:4]
mcmc.samples <- as.mcmc(post.samples)

### Gelman and Rubin
# Separate posterior samples into separate chains and save as mcmc.list object
mcmc.chain.1 <- as.mcmc(post.samples[1:2500,])
mcmc.chain.2 <- as.mcmc(post.samples[2501:5000,])
mcmc.chain.3 <- as.mcmc(post.samples[5001:7500,])
mcmc.chain.4 <- as.mcmc(post.samples[7501:10000,])
mcmc.chains.list <- as.mcmc.list( list(mcmc.chain.1, mcmc.chain.2,
                                         mcmc.chain.3, mcmc.chain.4) )
```

```
> gelman.diag(mcmc.chains.list)
Potential scale reduction factors:
```

	Point est.	Upper C.I.
beta0	1.01	1.02
beta1	1.01	1.02
beta2	1.00	1.00
sigma2	1.00	1.00

Multivariate psrf

```
### Geweke
> geweke.diag(mcmc.samples)

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

beta0   beta1   beta2   sigma2
1.0150 -0.4933 -1.4461 -2.0654
```

```
### Raftery and Lewis
> raftery.diag(mcmc.samples)

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95
```

	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
beta0	8	8856	3746	2.36
beta1	8	9476	3746	2.53
beta2	6	7066	3746	1.89
sigma2	5	6302	3746	1.68

```
### Heidelberger and Welch
> heidel.diag(mcmc.samples)
```

	Stationarity	start	p-value
	test	iteration	
beta0	passed	1	0.2405
beta1	passed	1	0.2699
beta2	passed	1	0.7276
sigma2	passed	1	0.0818

	Halfwidth	Mean	Halfwidth
	test		
beta0	passed	53.284	0.16838
beta1	passed	0.127	0.00123
beta2	passed	5.900	0.02020
sigma2	passed	7.372	0.11405

```
### Autocorrelations and cross-correlations
> round( autocorr.diag(mcmc.samples), digits = 4 )
      beta0   beta1   beta2   sigma2
Lag 0  1.0000  1.0000  1.0000  1.0000
Lag 1  0.4602  0.4326  0.3022  0.3867
Lag 5  0.0290  0.0195 -0.0035  0.0413
Lag 10 -0.0083 -0.0162 -0.0054 -0.0010
Lag 50  0.0065  0.0095 -0.0221 -0.0151

> round( crosscorr(mcmc.samples), digits = 4 )
      beta0   beta1   beta2   sigma2
beta0  1.0000 -0.8578 -0.4059 -0.0545
beta1 -0.8578  1.0000 -0.1011  0.0562
beta2 -0.4059 -0.1011  1.0000  0.0091
sigma2 -0.0545  0.0562  0.0091  1.0000
```

References

- 1) Cowles and Carlin (1996, JASA, pp-883-904) (excellent review article on Markov Chain Monte Carlo convergence diagnostics).
- 2) Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-72.
- 3) Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith). Clarendon Press, Oxford, UK.
- 4) Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-44.
- 5) Raftery, A. L. and Lewis, S. (1992a). Commnet: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7, 493-7
- 6) Raftery, A. L. and Lewis, S (1992b). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, (ed J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp 763-74. Oxford University Press.

Chapter 4:

Bayesian Analysis of Generalized Linear Models

Introduction to Generalized Linear Models

The class of generalized linear models is a natural generalization of the classical linear model. Generalized linear models include as special cases, linear regression and analysis of variance models, logit and probit models for quantal response data, log-linear models and multinomial response models for counts, and some commonly used models for survival data.

To simplify the transition from the classical normal linear model ($Y = X\beta + \epsilon$, $\epsilon \sim N_n(0, \sigma^2 I)$) to generalized linear models, it will be important to characterize specific aspects of the **linear** model:

1. **Random component:** $Y \sim N_n(\mu, \sigma^2 I)$, where $\mu = X\beta$. Note that the linear model has constant variance.
2. **Systematic component:** The **covariate** comprises the systematic component of the model. For the i^{th} observation, we let

$$\eta_i = x_i' \beta \quad , \quad i = 1, \dots, n .$$

We call η_i the **linear predictor**.

Thus $y_i \sim N(x'_i \beta, \sigma^2) = N(\eta_i, \sigma^2)$, $i = 1, \dots, n$, and the y_i 's are independent. Note here that for the usual normal linear model, the relationship between the **mean** of y_i and η_i is given by

$$\mu_i \equiv E(y_i | x_i, \beta) = x'_i \beta = \eta_i, \quad i = 1, \dots, n.$$

Thus

$$\mu_i = \eta_i.$$

Generalized linear models involve **2 extensions** of the normal linear model.

1. The distribution of y is from the **exponential family**.
2. The relationship between $\mu_i = E(y_i | x_i, \beta)$ can be made more general, so that

$$g(\mu_i) = \eta_i \equiv x'_i \beta.$$

$g(\mu_i)$ is called the **μ -link** function and relates the **mean of y_i** (i.e., μ_i) to the linear predictor η_i .

y has a distribution in the exponential family with **canonical parameter** θ and dispersion parameter ϕ if y has density

$$p(y | \theta, \phi) = \exp \{ [y\theta - b(\theta)]/\phi + c(y, \phi) \}. \quad (4.1)$$

Without loss of generality, we assume $a(\phi) = \phi$, so that

$$p(y | \theta, \phi) = \exp \{ [y\theta - b(\theta)]/\phi + c(y, \phi) \} .$$

Here

$$\int_y \exp \{ (y\theta - b(\theta))/\phi + c(y, \phi) \} dy = 1 ,$$

so that

$$\exp \left\{ \frac{b(\theta)}{\phi} \right\} = \int_y \exp \left\{ \frac{y\theta}{\phi} + c(y, \phi) \right\} dy .$$

Here $b(\cdot)$ and $c(\cdot)$ are **known** functions. If ϕ is unknown, (4.1) may or may **not** be an exponential family. θ is called the **canonical parameter**. An excellent book on generalized linear models is McCullagh & Nelder (1989, Chapman and Hall).

The class of generalized linear models has many uses in biostatistics. Binomial models are often used to model dose response. Gamma models are often used to model survival or time-to-event data. Poisson models are used to model count data, such as yearly pollen counts, number of cancerous nodes, etc.

Distributions included in the exponential family are the normal, binomial, gamma, Poisson, beta, multinomial, and inverse Gaussian distributions. To see how the normal distribution, for example, fits into the framework above, suppose

$$y \sim N(\mu, \sigma^2) .$$

Then

$$\begin{aligned} & p(y | \mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{(y\mu - \frac{\mu^2}{2})/\sigma^2 - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\} , \end{aligned}$$

so that in this case,

$$\theta = \mu$$

$$a(\phi) \equiv \phi = \sigma^2$$

$$b(\theta) = \frac{\theta^2}{2}$$

$$c(y, \phi) = -\frac{1}{2} \left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right] .$$

As a second example, suppose

$$y \sim \text{Binomial}(N, p) .$$

Then,

$$\begin{aligned} p(y | N, p) &= \binom{N}{y} p^y (1-p)^{N-y} \\ &= \exp \left\{ y \log(p) + (N-y) \log(1-p) + \log \binom{N}{y} \right\} \\ &= \exp \left\{ y \log \left(\frac{p}{1-p} \right) - N \log \left(\frac{1}{1-p} \right) + \log \binom{N}{y} \right\} \\ &= \exp \left\{ N \left[\frac{y}{N} \log \left(\frac{p}{1-p} \right) - \log \left(\frac{1}{1-p} \right) \right] + \log \binom{N}{y} \right\} \end{aligned}$$

so that

$$\begin{aligned}
 \theta &= \log\left(\frac{p}{1-p}\right) \\
 \Rightarrow p &= \frac{\exp\{\theta\}}{1+\exp\{\theta\}} \\
 \phi &= \frac{1}{N} \\
 c(y, \phi) &= \log\left(\frac{N}{y}\right) \\
 b(\theta) &= \log\left(\frac{1}{1-p}\right) \\
 &= -\log(1-p) \\
 &= -\log\left(1 - \frac{\exp\{\theta\}}{1+\exp\{\theta\}}\right) \\
 &= -\log\left(\frac{1}{1+\exp\{\theta\}}\right) \\
 &= \log(1 + \exp\{\theta\}) .
 \end{aligned}$$

Thus $b(\theta) = \log(1 + \exp\{\theta\})$. The transformation $\log\left(\frac{p}{1-p}\right)$ is sometimes called the **logit transformation**.

Tables 4.1 and 4.2 show the component of an exponential family for several specific distributions.

Table 4.1: Characteristics of some common univariate distributions in the exponential family[†]

	Normal	Poisson	Binomial
Notation	$N(\mu, \sigma^2)$	$\text{Poisson}(\mu)$	$\text{Binomial}(m, \pi)/m$
Range of y	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$
Dispersion parameter: ϕ	$\phi = \sigma^2$	1	$1/m$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp\{\theta\}$	$\log(1 + \exp\{\theta\})$
$c(y; \phi)$	$-\frac{1}{2} \left(\frac{y^2}{\phi} + \log(2\pi\phi) \right)$	$-\log y!$	$\log \binom{m}{my}$
$\mu(\theta) = E(Y; \theta)$	ϕ	$\exp\{\theta\}$	$\frac{\exp\{\theta\}}{1+\exp\{\theta\}}$
Canonical link: $\theta(\mu)$	identity	log	logit
Variance function: $V(\mu)$	1	μ	$\mu(1 - \mu)$

[†]The mean-value parameter is denoted by μ , or by π for the binomial distribution.

Table 4.2: Characteristics of some common univariate distributions in the exponential family
 (continued)[†]

	Gamma	Inverse Gaussian
Notation	$\text{gamma}(\mu, \nu)$	$IG(\mu, \sigma^2)$
Range of y	$(0, \infty)$	$(0, \infty)$
Dispersion parameter: ϕ	$\phi = \nu^{-1}$	$\phi = \sigma^2$
Cumulant function: $b(\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y; \phi)$	$\nu \log(\nu y) - \log y - \log \Gamma(\nu)$	$-\frac{1}{2} \left\{ \log \left(2\pi\phi y^3 \right) + \frac{1}{\phi y} \right\}$
$\mu(\theta) = E(Y; \theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
Canonical link: $\theta(\mu)$	reciprocal	$1/\mu^2$
Variance function: $V(\mu)$	μ^2	μ^3

[†]The mean-value parameter is denoted by μ .

The parameterization of the gamma distribution is such that its variance is μ^2/ν .

Let us denote the log-likelihood for one observation y by

$$\begin{aligned} l(\theta, \phi) &= \log [p(y | \theta, \phi)] \\ &= \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \end{aligned}$$

For exponential family models, we have the following two identities:

$$\begin{aligned} E\left[\frac{\partial l(\theta, \phi)}{\partial \theta}\right] &= 0, \\ E\left[\frac{\partial^2 l(\theta, \phi)}{\partial \theta^2}\right] + E\left(\frac{\partial l(\theta, \phi)}{\partial \theta}\right)^2 &= 0. \end{aligned}$$

These two identities lead to some properties concerning the mean of y , i.e., $E(y | \theta, \phi)$ and the $\text{Var}(y | \theta, \phi)$.

$$\begin{aligned} \frac{\partial l(\theta, \phi)}{\partial \theta} &= \frac{y - b'(\theta)}{\phi}, \quad b'(\theta) = \frac{db(\theta)}{d\theta} \\ \frac{\partial^2 l(\theta, \phi)}{\partial \theta^2} &= -\frac{b''(\theta)}{\phi}, \quad b''(\theta) = \frac{d^2 b(\theta)}{d\theta^2}. \end{aligned}$$

From the identities above, we have

$$\begin{aligned} 0 = E \left[\frac{\partial l(\theta, \phi)}{\partial \theta} \right] &= \phi^{-1} E(y | \theta, \beta) - b'(\theta) \\ &= \phi^{-1} (\mu - b'(\theta)) , \end{aligned}$$

so that

$$E(y | \theta, \phi) \equiv \mu = b'(\theta) .$$

Thus

$$\mu = b'(\theta) .$$

Similarly, using the second identity, we have

$$\begin{aligned} 0 &= -\frac{b''(\theta)}{\phi} + \frac{\text{Var}(y | \theta, \phi)}{\phi^2} \\ \Rightarrow \quad \text{Var}(y | \theta, \phi) &= \phi b''(\theta) . \end{aligned}$$

This tells us that once we know the $b(\cdot)$ function, we can immediately compute the mean and variance of the exponential family model.

Now suppose we have n independent observations y_1, \dots, y_n from an exponential family as in (4.1).

Then the density for the i^{th} observation can be written as

$$p(y_i | \theta_i, \phi) = \exp \left\{ \phi^{-1} (y_i \theta_i - b(\theta_i)) + c(y_i, \phi) \right\} .$$

The density based on n observations is

$$p(y | \theta, \phi) = \prod_{i=1}^n p(y_i | \theta_i, \phi) ,$$

where $y = (y_1, \dots, y_n)$, $\theta = (\theta_1, \dots, \theta_n)$.

To construct the regression model, (i.e., the generalized linear model), we let the θ_i 's depend on the linear predictor $\eta_i = x'_i \beta$ through the equation

$$\theta_i = \theta(\eta_i) , \quad i = 1, \dots, n , \tag{4.2}$$

where $x'_i = (x_{i1}, \dots, x_{ip})$ and $\beta = (\beta_1, \dots, \beta_p)'$.

The link function in (4.2) is called the **θ -link** and is often more convenient to use than the μ -link. The θ -link is a one-to-one function of the μ -link.

Once $\theta_i = \theta(\eta_i)$ is given, one can write the likelihood function as a function in (β, ϕ) . When $\theta_i = \eta_i$, we say that we have a **canonical link**. The function $\theta_i = \theta(\eta_i)$ can be any **monotonic** function.

Example 4.1

Suppose $y_i \sim \text{Binomial}(1, p_i)$, the y_i 's are independent, $i = 1, \dots, n$. We have

$$\begin{aligned} p(y_i | p_i) &= \exp \left\{ y_i \log \left(\frac{p_i}{1-p_i} \right) - \log \left(\frac{1}{1-p_i} \right) \right\} \\ &= \exp [y_i \theta_i - \log(1 + \exp \{\theta_i\})] . \end{aligned}$$

If a canonical link is used, then we set $\theta_i = \eta_i = x_i' \beta$. Substituting $\theta_i = x_i' \beta$ into $p(y_i | p_i)$ above, we get

$$p(y_i | \beta) = \exp [y_i x_i' \beta - \log(1 + \exp \{x_i' \beta\})] .$$

Thus, the likelihood function of β based on all n observations is given by

$$\begin{aligned}
 p(y | \beta) &= \prod_{i=1}^n p(y_i | \beta) \\
 &= \prod_{i=1}^n \exp [y_i x'_i \beta - \log(1 + \exp \{x'_i \beta\})] \\
 &= \exp \left[\sum_{i=1}^n \{y_i x'_i \beta - \log(1 + \exp \{x'_i \beta\})\} \right] .
 \end{aligned}$$

For this model, the relation between θ_i and μ_i is

$$\theta_i = \log \left(\frac{\mu_i}{1-\mu_i} \right), \quad \text{where } \mu_i = E(y_i | p_i) \equiv p_i .$$

Thus

$$\mu_i = \frac{\exp \{\theta_i\}}{1 + \exp \{\theta_i\}} .$$

Suppose, we consider a **probit model**. The μ -link for the probit model is given by

$$\Phi^{-1}(\mu_i) = x'_i \beta \equiv \eta_i .$$

That is, $g(\mu_i) = \Phi^{-1}(\mu_i)$ is the **μ -link**, $i = 1, \dots, n$. Since $\mu_i = \frac{\exp\{\theta_i\}}{1+\exp\{\theta_i\}}$, we can solve for the θ -link.

We have

$$\begin{aligned}\Phi^{-1}(\mu_i) &= \eta_i \\ \Rightarrow \mu_i &= \Phi(\eta_i), \quad \text{where } \eta_i = x'_i \beta, \\ \Rightarrow \Phi(\eta_i) &= \frac{\exp\{\theta_i\}}{1+\exp\{\theta_i\}}.\end{aligned}$$

Thus,

$$\begin{aligned}\theta_i &= \log \left(\frac{\Phi(\eta_i)}{1 - \Phi(\eta_i)} \right) \\ &= \log \left(\frac{\Phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right).\end{aligned}$$

Thus the likelihood function based on n observations for the probit model is given by

$$\begin{aligned}
 p(y | \beta) &= \prod_{i=1}^n \exp [y_i \theta_i - \log(1 + \exp \{\theta_i\})] \\
 &= \prod_{i=1}^n \exp \left\{ y_i \log \left(\frac{\Phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right) - \log \left(\frac{1}{1 - \Phi(x'_i \beta)} \right) \right\} \\
 &= \exp \left\{ \sum_{i=1}^n \left[y_i \log \left(\frac{\Phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right) - \log \left(\frac{1}{1 - \Phi(x'_i \beta)} \right) \right] \right\} .
 \end{aligned}$$

Exercise 4.1

Find the likelihood function of β for the Poisson model based on n observation,

- i) with a canonical link
- ii) with the link $\theta_i = \log(x'_i \beta)$.

Any model that satisfies

$$p(y_i \mid \theta_i, \phi) = \exp \left\{ \phi^{-1}(y_i \theta_i - b(\theta_i)) + c(y_i, \phi) \right\}$$

and

$$\theta_i = \theta(\eta_i), \quad \eta_i = x'_i \beta,$$

is called a generalized linear model (GLM).

Below we give some distributions with their canonical links.

<u>Distribution</u>	<u>Canonical μ-link, where $\mu = E(y \mid \mu)$</u>
Normal	$\eta = \mu$
Poisson	$\eta = \log(\mu)$
Binomial	$\eta = \log \left(\frac{\mu}{1-\mu} \right)$
Gamma	$\eta = \mu^{-1}$

Estimation in the Generalized Linear Model

The maximum likelihood estimate of β **does not** have a closed analytic form in general for GLM's. We have to rely on iterative methods such as Newton-Raphson or Fisher scoring to obtain MLE's.

Let us first derive the likelihood equations for β . We have, for an arbitrary GLM,

$$\begin{aligned} l(\beta, \phi) &= \sum_{i=1}^n \left\{ \phi^{-1} [y_i \theta_i - b(\theta_i)] + c(y_i, \phi) \right\} \\ &= \sum_{i=1}^n l_i(\theta_i, \phi) \equiv \sum_{i=1}^n l_i . \end{aligned}$$

Note that in the above equation, θ_i depends on β through $\theta_i = \theta(x'_i \beta) \equiv \theta(\eta_i)$.

Also note that for **ANY** GLM, the likelihood function depends on β **only** through $x'_i \beta = \eta_i$. Moreover, note that $\phi \equiv 1$ for the Bernoulli and Poisson models.

Now

$$\begin{aligned}
 \frac{\partial l(\beta, \phi)}{\partial \beta_j} &= \sum_{i=1}^n \frac{\partial}{\partial \beta_j} l_i(\theta_i, \phi) \\
 &\equiv \sum_{i=1}^n \frac{\partial}{\partial \beta_j} l_i(\theta(x'_i \beta), \phi) \\
 &= \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\
 &= \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} x_{ij} , \quad j = 1, \dots, p .
 \end{aligned}$$

Note that since $\eta_i = x'_i \beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p$, we have

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} .$$

Let $\Delta_i = \frac{\partial \theta_i}{\partial \eta_i}$. Δ_i is sometimes called the link adjustment. If a canonical link is used, then $\Delta_i = 1$, since in this case, $\theta_i = \eta_i$.

We can now write

$$\begin{aligned}\frac{\partial l_i}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} (\phi^{-1} [y_i \theta_i - b(\theta_i)] + c(y_i, \phi)) \\ &= \frac{y_i - b'(\theta_i)}{\phi} \\ &= \frac{y_i - \mu_i}{\phi}.\end{aligned}$$

Thus, the likelihood equations for β can be written as

$$\frac{\partial}{\partial \beta_j} l(\beta, \phi) = 0$$

$$\Leftrightarrow \sum_{i=1}^n (y_i - \mu_i) \Delta_i x_{ij} = 0, \quad j = 1 \dots, p.$$

In matrix form, we can write the likelihood equations as

$$X' \Delta (y - \mu) = 0 ,$$

where

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & & x_{np} \end{bmatrix}_{n \times p}, \quad \Delta = \begin{pmatrix} \Delta_1 & & 0 \\ & \ddots & \\ 0 & & \Delta_n \end{pmatrix}_{n \times n},$$

$$y = (y_1, \dots, y_n)', \text{ and } \mu = (\mu_1, \dots, \mu_n)' .$$

Letting $S = y - \mu$, the equations become

$$X' \Delta S = 0 . \tag{4.3}$$

We note that $\Delta = I$ under a **canonical** link. The equations in (4.3) are **non-linear** in β . We can carry out a Newton-Raphson procedure for obtaining the MLE of β , denoted by $\hat{\beta}$.

Note that the likelihood equations above for β do **NOT** depend on ϕ , and thus estimation of β is **independent** of ϕ .

The Newton-Raphson iterative procedure for β can be written as

$$\beta^{(t+1)} = \beta^{(t)} - \left[\frac{\partial}{\partial \beta} (X' \Delta S) \right]^{-1} (X' \Delta S) \Big|_{\beta=\beta^{(t)}} .$$

It can be shown that

$$\frac{\partial}{\partial \beta} (X' \Delta S) = -X'(\Delta V \Delta - \dot{\Delta} H)X ,$$

where

$$H = \begin{pmatrix} y_1 - \mu_1 & 0 \\ & \ddots \\ 0 & y_n - \mu_n \end{pmatrix}_{n \times n} ,$$

$$V = \begin{pmatrix} b''(\theta_1) & 0 \\ & \ddots \\ 0 & b''(\theta_n) \end{pmatrix}_{n \times n} ,$$

$b''(\theta_i) = \text{Var}(y_i | \theta_i)$, and $\theta_i = \theta(x'_i \beta)$. V is called the matrix of **variance functions**. Also, we have

$$\dot{\Delta} = \begin{pmatrix} \frac{\partial^2 \theta_1}{\partial \eta_1^2} & 0 \\ & \ddots \\ 0 & \frac{\partial^2 \theta_n}{\partial \eta_n^2} \end{pmatrix}_{n \times n}.$$

Also note that

$$\frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \beta'} = -\frac{1}{\phi} X' (\Delta V \Delta - \dot{\Delta} H) X .$$

We can now rewrite the Newton-Raphson iterative procedure as

$$\beta^{(t+1)} = \beta^{(t)} + [X' (\Delta V \Delta - \dot{\Delta} H) X]^{-1} (X' \Delta S) |_{\beta=\beta^{(t)}} .$$

Under a **canonical link**, $\Delta = I$, $\dot{\Delta} = 0$, so that the algorithm reduces to

$$\beta^{(t+1)} = \beta^{(t)} + (X' V X)^{-1} (X' S) |_{\beta=\beta^{(t)}} .$$

It can be shown that under **canonical links**, the Newton-Raphson iterative scheme above can be expressed as an **iteratively reweighted least squares** (IRLS) algorithm. To see this, define

$$Z^{(t)} = X\beta^{(t)} + V^{-1}S \Big|_{\beta=\beta^{(t)}} .$$

Since

$$\beta^{(t)} = (X'VX)^{-1}(X'VX)\beta^{(t)}$$

and

$$(X'VX)^{-1}(X'S) = (X'VX)^{-1}X'(VV^{-1})S ,$$

we can write

$$\begin{aligned}\beta^{(t+1)} &= (X'VX)^{-1}(X'V)Z^{(t)} \Big|_{\beta=\beta^{(t)}} \\ &= (X'V^{(t)}X)^{-1}X'V^{(t)}Z^{(t)} ,\end{aligned}$$

where $V^{(t)} = V \Big|_{\beta=\beta^{(t)}}$. So we see that $\beta^{(t+1)}$ corresponds to a weighted least squares regression of $Z^{(t)}$ on X with weight matrix $V^{(t)}$.

Fisher Scoring

In Fisher scoring, we use the **expected Hessian** instead of the Hessian in the Newton-Raphson algorithm for β . The Hessian is given by

$$\text{Hess}(\beta) = -X'(\Delta V \Delta - \dot{\Delta} H)X .$$

$$\begin{aligned} E_{y|\beta}[-X'(\Delta V \Delta - \dot{\Delta} H)X] \\ = -X'(\Delta V \Delta - \dot{\Delta} E(H))X \\ = -X'\Delta V \Delta X . \end{aligned}$$

$$\text{Since } H = \begin{pmatrix} y_1 - \mu_1 & & 0 \\ & \ddots & \\ 0 & & y_n - \mu_n \end{pmatrix} ,$$

$$\text{we have } E(H) = 0 .$$

Note that the **Fisher Information matrix** of β is given by

$$I(\beta) = -\frac{1}{\phi}(-X'\Delta V \Delta X) = \frac{1}{\phi}X'\Delta V \Delta X .$$

Thus the Fisher scoring algorithm is

$$\beta^{(t+1)} = \beta^{(t)} + (X' \Delta V \Delta X)^{-1} X' \Delta S \Big|_{\beta=\beta^{(t)}} .$$

When a **canonical link** is used

$$\text{Newton-Raphson} = \text{Fisher scoring} = \text{IRLS} .$$

Example 4.2: Logistic Regression

For logistic regression, $\phi = 1$ and we have a canonical link. The log-likelihood function of β is given by

$$l(\beta) = \sum_{i=1}^n \{y_i x'_i \beta - \log(1 + \exp\{x'_i \beta\})\} .$$

Let us set up the Newton-Raphson iterative scheme for β . We have

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} + (X' V X)^{-1} X' S \Big|_{\beta=\beta^{(t)}} , \\ S &= \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix} = \begin{pmatrix} y_1 - \frac{\exp\{x'_1 \beta\}}{1+\exp\{x'_1 \beta\}} \\ \vdots \\ y_n - \frac{\exp\{x'_n \beta\}}{1+\exp\{x'_n \beta\}} \end{pmatrix} , \end{aligned}$$

$$\mu_i = \frac{\exp\{\theta_i\}}{1 + \exp\{\theta_i\}} = \frac{\exp\{x'_i \beta\}}{1 + \exp\{x'_i \beta\}} , \quad i = 1, \dots, n .$$

Now

$$V = \begin{pmatrix} b''(\theta_1) & & 0 \\ & \ddots & \\ & & b''(\theta_n) \end{pmatrix} , \quad b(\theta_i) = \log(1 + \exp\{\theta_i\})$$

$$\begin{aligned} b''(\theta_i) &= \frac{\exp\{\theta_i\}}{(1+\exp\{\theta_i\})^2} \\ &= \frac{\exp\{x'_i \beta\}}{(1+\exp\{x'_i \beta\})^2} . \end{aligned}$$

Thus

$$V = \begin{pmatrix} \frac{\exp\{x'_1 \beta\}}{(1+\exp\{x'_1 \beta\})^2} & & 0 \\ & \ddots & \\ & & \frac{\exp\{x'_n \beta\}}{(1+\exp\{x'_n \beta\})^2} \end{pmatrix}_{n \times n} .$$

Note that since we are using a canonical link, $\Delta = I$ and $\dot{\Delta} = 0$.

Frequentist Inference for GLM's

The sampling distributions of the estimates, test statistics, or interval estimates do not have analytic closed forms in general. Thus we have to rely on asymptotic theory ($n \rightarrow \infty$) to do inference for GLM's.

If we expand the likelihood function $L(\beta)$ in a Taylor's series about $\hat{\beta}$, we get

$$L(\beta) \approx \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta})' (X' \Delta V \Delta X) (\beta - \hat{\beta}) \right\} .$$

Thus, we are led to the following theorem.

Theorem 4.1

Assume $\phi = 1$. For noncanonical links, as $n \rightarrow \infty$

$$\hat{\beta} \rightarrow N_p(\beta, (X' \Delta V \Delta X)^{-1}) ,$$

and for canonical links

$$\hat{\beta} \rightarrow N_p(\beta, (X' V X)^{-1}) .$$

We make use of this asymptotic result to do hypothesis testing, interval estimation, model selection, etc.

Hypothesis testing

Suppose we have two nested models $m \subset m_0$. Then

$$-2 \log \lambda \rightarrow \chi^2_{(k_{m_0} - k_m)} ,$$

where

$$\lambda = \frac{p(y | \hat{\beta}, m)}{p(y | \hat{\beta}, m_0)}$$

is the likelihood ratio statistic for testing m against m_0 and $k_{m_0} = \text{dimension}(m_0)$, $k_m = \text{dimension}(m)$.

The goodness of fit of a GLM can be assessed by computing the deviance. For a given model m , the deviance is defined as

$$D = -2[l(\hat{\beta}^{(m)}) - l(y)] ,$$

where $l(\cdot)$ is the log-likelihood function. We have

$$D \rightarrow \chi^2_{n-k_m} .$$

The deviance for various GLM's is given below:

$$\text{Normal: } D = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 , \quad \mu_i = E(y_i | \mu_i)$$

$$\text{Poisson: } D = 2 \sum_{i=1}^n \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right\}$$

$$\text{Binomial: } y_i \sim \text{Binomial}(N_i, p_i) , \quad \mu_i = N_i p_i$$

$$D = 2 \sum_{i=1}^n \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (N_i - y_i) \log \left(\frac{N_i - y_i}{N_i - \hat{\mu}_i} \right) \right\}$$

$$\text{Gamma: } D = 2 \sum_{i=1}^n \left\{ -\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right\}$$

Note that for the **canonical links**

$$\text{Normal: } \hat{\mu}_i = x_i' \hat{\beta}$$

$$\text{Poisson: } \hat{\mu}_i = \exp \left\{ x_i' \hat{\beta} \right\}$$

$$\text{Binomial: } \hat{\mu}_i = \frac{N_i \exp \{ x_i' \hat{\beta} \}}{1 + \exp \{ x_i' \hat{\beta} \}}$$

$$\text{Gamma: } \hat{\mu}_i = \frac{1}{x_i' \hat{\beta}} .$$

Prior Distributions for Generalized Linear Models

Let us first consider classes of priors for exponential family models. Let us assume $\phi = 1$ throughout, since this is the case for the binomial, Poisson, and exponential distributions. The exponential family density can be written as

$$p(y | \theta) = \exp\{y\theta - b(\theta) + c(y)\} . \quad (4.4)$$

We can develop a class of conjugate priors for the exponential family. We are led to the following theorem.

Theorem 4.2

Suppose $p(y | \theta)$ is given by (4.4). Suppose we consider a prior on θ of the form

$$\pi(\theta) \propto \exp\{a_0(\theta y_0 - b(\theta))\} , \quad (4.5)$$

where (a_0, y_0) are specified hyperparameters. Then $\pi(\theta)$ is a conjugate prior for θ , which indexes the exponential family $p(y | \theta)$. (Note: Conjugate priors are proper priors by definition.)

Proof:

Label the prior distribution in (4.5) by $D(a_0, y_0)$. That is, a priori

$$\theta \sim D(a_0, y_0) \Rightarrow \pi(\theta) \propto \exp \{a_0(\theta y_0 - b(\theta))\} .$$

We want to show that $\theta | y$ is also of type D .

$$\begin{aligned} p(\theta | y) &\propto p(y | \theta) \pi(\theta) \\ &\propto \exp \{y\theta - b(\theta)\} \exp \{a_0(\theta y_0 - b(\theta))\} \\ &= \exp \{\theta(y + a_0 y_0) - (a_0 + 1) b(\theta)\} \\ &= \exp \left\{ (a_0 + 1) \left[\theta \left(\frac{y + a_0 y_0}{a_0 + 1} \right) - b(\theta) \right] \right\} \end{aligned}$$

Thus $\theta | y \sim D\left(a_0 + 1, \frac{y + a_0 y_0}{a_0 + 1}\right)$.

Thus $\pi(\theta)$ in (4.5) is a **conjugate** prior for θ for the exponential family $p(y | \theta)$ given in (4.4). Note that a posteriori, the hyperparameters are $\left(a_0 + 1, \frac{y + a_0 y_0}{a_0 + 1}\right)$.

The quantity $\frac{y+a_0y_0}{a_0+1} = \left(\frac{1}{a_0+1}\right)y + \left(\frac{a_0}{a_0+1}\right)y_0$ is a weighted average of the data and the prior parameter y_0 , with the weights being defined by a_0 . Also the parameter a_0 is updated to $a_0 + 1$ in going from prior to posterior, after the collection of 1 observation y .

If we have n i.i.d. observations y_1, \dots, y_n , with

$$p(y_i | \theta) = \exp\{y_i\theta - b(\theta) + c(y_i)\}$$

and

$$\pi(\theta) \propto \exp\{a_0(y_0\theta - b(\theta))\},$$

then

$$\begin{aligned} p(\theta | y) &\propto \prod_{i=1}^n p(y_i | \theta) \pi(\theta) \\ &\propto \exp\{\theta \sum y_i - n b(\theta)\} \exp\{a_0(\theta y_0 - b(\theta))\} \\ &\propto \exp\left\{(n + a_0) \left(\theta \left(\frac{n\bar{y} + a_0 y_0}{n + a_0}\right) - b(\theta)\right)\right\}. \end{aligned}$$

Thus

$$\theta | y \sim D\left(n + a_0, \frac{n\bar{y} + a_0 y_0}{n + a_0}\right).$$

Theorem 4.3

Suppose we have n i.i.d. observations y_1, \dots, y_n , with

$$p(y_i | \theta) = \exp\{y_i\theta - b(\theta) + c(y_i)\}.$$

Let y_{0i} be a prior prediction for the marginal mean $E(y_i)$, $i = 1, \dots, n$. Chen and Ibrahim (2003, *Statistica Sinica*) show that a generalization of the prior on a previous page is

$$\begin{aligned}\pi(\theta) &\propto \prod_{i=1}^n \exp\{a_0(y_{0i}\theta - b(\theta))\} \\ &\propto \exp\{a_0 n \bar{y}_0 \theta - a_0 n b(\theta)\} \\ &= \exp\{a_0 n (\theta \bar{y}_0 - b(\theta))\}\end{aligned}$$

so that

$$\begin{aligned}p(\theta|y) &\propto \exp\{\theta n \bar{y} - nb(\theta)\} \exp\{a_0 n (\theta \bar{y}_0 - b(\theta))\} \\ &\propto \exp\left\{(n + a_0 n) \left(\theta \left(\frac{n \bar{y} + a_0 n \bar{y}_0}{n + a_0 n}\right) - b(\theta)\right)\right\}.\end{aligned}$$

Thus

$$\theta|y \sim D\left(n + a_0 n, \frac{\bar{y} + a_0 \bar{y}_0}{1 + a_0}\right).$$

Example 4.3: Normal Distribution

i) Suppose $y \sim N(\mu, 1)$. In this case, we have $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$. Thus

$$\begin{aligned} p(y | \theta) &\propto \exp \left\{ -\frac{1}{2} (y - \theta)^2 \right\} \\ \pi(\theta) &\propto \exp \{a_0(\theta y_0 - b(\theta))\} \\ &= \exp \left\{ a_0 \left(\theta y_0 - \frac{\theta^2}{2} \right) \right\} \\ &\propto \exp \left\{ -\frac{a_0}{2} (\theta - y_0)^2 \right\}. \end{aligned}$$

Thus $\theta \sim N(y_0, a_0^{-1})$ is the conjugate prior, where
 $-\infty < y_0 < \infty$, $a_0 > 0$.

- ii) Suppose we have y_1, \dots, y_n i.i.d. $N(\theta, 1)$. In this case, we have $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$. Thus

$$\begin{aligned}
 p(y|\theta) &\propto \exp\left\{-\frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2\right\} \\
 \pi(\theta) &\propto \exp\{a_0(\theta n \bar{y}_0 - nb(\theta))\}, \quad \bar{y}_0 = \frac{1}{n} \sum_{i=1}^n y_{0i} \\
 &\propto \exp\left\{a_0 \left(\theta n \bar{y}_0 - n \frac{\theta^2}{2}\right)\right\} \\
 &\propto \exp\left\{-\frac{a_0 n}{2} (\theta - \bar{y}_0)^2\right\}
 \end{aligned}$$

Thus $\theta \sim N(\bar{y}_0, (a_0 n)^{-1})$ is the conjugate prior, where $-\infty < y_0 < \infty$, $a_0 > 0$.

Example 4.4: Binomial Distribution

i) Suppose $y \sim \text{binomial}(1, p)$. In this case, we have

$$p = \frac{\exp\{\theta\}}{1+\exp\{\theta\}}, \quad b(\theta) = \log(1 + \exp\{\theta\}). \quad \text{Thus}$$

$$p(y | \theta) \propto \exp\{y\theta - \log(1 + \exp\{\theta\})\}$$

$$\pi(\theta) \propto \exp\{a_0(\theta y_0 - b(\theta))\}$$

$$= \exp\{a_0(\theta y_0 - \log(1 + \exp\{\theta\}))\}$$

$$= \frac{\exp\{a_0 y_0 \theta\}}{(1 + \exp\{\theta\})^{a_0}}$$

$$= \left(\frac{\exp\{\theta\}}{1 + \exp\{\theta\}} \right)^{a_0 y_0} \left(\frac{1}{1 + \exp\{\theta\}} \right)^{a_0 (1 - y_0)}$$

Thus

$$\pi(\theta) \propto \left(\frac{\exp\{\theta\}}{1 + \exp\{\theta\}} \right)^{a_0 y_0} \left(\frac{1}{1 + \exp\{\theta\}} \right)^{a_0 (1 - y_0)}.$$

This prior is equivalent to a **beta prior** for θ . To see this, let $p = \frac{\exp\{\theta\}}{1+\exp\{\theta\}}$, $1-p = \frac{1}{1+\exp\{\theta\}}$. Then

$$\frac{dp}{d\theta} = \frac{\exp\{\theta\}}{(1 + \exp\{\theta\})^2} = p(1 - p) .$$

Then

$$\begin{aligned}\pi(p) &\propto p^{a_0 y_0} (1-p)^{a_0(1-y_0)} (p^{-1}(1-p)^{-1}) \\ &= p^{a_0 y_0 - 1} (1-p)^{a_0(1-y_0)-1},\end{aligned}$$

and thus $p \sim \text{beta}(a_0 y_0, a_0(1 - y_0))$ where $p = \frac{\exp\{\theta\}}{1+\exp\{\theta\}}$, $a_0 > 0$, $0 < y_0 < 1$.

Thus, the conjugate prior for the binomial family is a beta prior.

ii) If y_1, \dots, y_n are i.i.d. binomial($1, p$), then

$$\pi(\theta) \propto \left(\frac{\exp\{\theta\}}{1 + \exp\{\theta\}} \right)^{a_0 n \bar{y}_0} \left(\frac{1}{1 + \exp\{\theta\}} \right)^{a_0 n (1 - \bar{y}_0)}.$$

If $p = \frac{\exp\{\theta\}}{1 + \exp\{\theta\}}$, then

$$p \sim \text{beta}(na_0 \bar{y}_0, na_0(1 - \bar{y}_0)).$$

Example 4.5: Poisson Distribution

- i) Suppose $y \sim \text{Poisson}(\mu)$. In this case, we have
 $\mu = \exp\{\theta\}$, $b(\theta) = \exp\{\theta\}$. Thus

$$\begin{aligned} p(y | \theta) &\propto \exp\{y\theta - \exp\{\theta\}\} \\ \pi(\theta) &\propto \exp\{a_0(\theta y_0 - b(\theta))\} \\ &= \exp\{a_0(\theta y_0 - \exp\{\theta\})\}. \end{aligned}$$

Thus $\pi(\theta) \propto \exp\{a_0(\theta y_0 - \exp\{\theta\})\}$. This prior is equivalent to a **gamma prior** for θ .

To see this, let $u = \exp\{\theta\}$. Then $\frac{du}{d\theta} = \exp\{\theta\} = u$, and

$$\begin{aligned} \pi(u) &\propto u^{a_0 y_0} \exp\{-a_0 u\} u^{-1} \\ &= u^{a_0 y_0 - 1} \exp\{-a_0 u\}. \end{aligned}$$

Thus $u \sim \text{gamma}(a_0 y_0, a_0)$, where $a_0 > 0$, $y_0 > 0$.

Therefore, the gamma prior is the conjugate prior for the Poisson model.

ii) Suppose y_1, \dots, y_n are i.i.d. Poisson(u). Then

$$\pi(\theta) \propto \exp \{na_0 (\theta \bar{y}_0 - \exp\{\theta\})\}$$

If $u = \exp\{\theta\}$, then

$$u \sim \text{gamma}(na_0 \bar{y}_0, na_0).$$

Example 4.6: Exponential Distribution

- i) Suppose $y \sim \text{exponential}(\lambda)$, so that $p(y | \lambda) = \lambda \exp\{-\lambda y\}$. Here we have $p(y | \lambda) = \exp\{y(-\lambda) - \log\left(\frac{1}{\lambda}\right)\}$. We see that $\theta = -\lambda$, and $b(\theta) = \log\left(\frac{1}{\lambda}\right) = \log\left(\frac{1}{-\theta}\right)$, $\theta < 0$.

Thus

$$\begin{aligned} p(y | \theta) &= \exp\left\{y\theta - \log\left(\frac{1}{-\theta}\right)\right\} \\ \pi(\theta) &\propto \exp\{a_0(\theta y_0 - b(\theta))\} \\ &= \exp\left\{a_0\left(\theta y_0 - \log\left(\frac{1}{-\theta}\right)\right)\right\} \\ &= \theta^{a_0} \exp\{a_0 y_0 \theta\}, \quad \theta < 0 \\ &= \theta^{a_0+1-1} \exp\{a_0 y_0 \theta\}, \quad \theta < 0. \end{aligned}$$

This prior is equivalent to a **gamma** prior. To see this, let $u = -\theta$, then

$$\pi(u) = u^{a_0+1-1} \exp\{-a_0 y_0 u\}, \quad u > 0.$$

Thus $u \sim \text{gamma}(a_0 + 1, a_0 y_0)$, $a_0 > 0$, $y_0 > 0$.

Thus the conjugate prior for the exponential distribution is the gamma prior. A similar result can be shown if $y \sim \text{gamma}(\alpha, \lambda)$, where α is known. That is, if $y \sim \text{gamma}(\alpha, \lambda)$, where α is known, then a conjugate prior for λ is a gamma prior. (Note that Example 4.9 uses $\alpha = 1$).

In this case, we have

$$\begin{aligned} p(y | \alpha, \lambda) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\lambda y\} \\ &= \exp \left\{ y(-\lambda) - \alpha \log \left(\frac{1}{\lambda} \right) + (\alpha - 1) \log(y) - \log(\Gamma(\alpha)) \right\}. \end{aligned}$$

Here $\phi = \frac{1}{\alpha}$ and $c(y, \phi) = (\alpha - 1) \log(y) - \log(\Gamma(\alpha))$.

Thus

$$p(y^* | \alpha, \lambda) = \exp \{ [(y^* \theta - b(\theta)) / \phi + c(y, \phi)] \}$$

where $y^* = y/\alpha$, $\theta = -\lambda$, $b(\theta) = \log \left(\frac{1}{-\theta} \right)$, $\phi = \alpha^{-1}$ (ϕ is known).

Thus

$$\pi(\theta) \propto \exp \left\{ a_0 \alpha \left(\theta y_0^* - \log \left(\frac{1}{-\theta} \right) \right) \right\},$$

which is equivalent to a $\text{gamma}(a_0 \alpha + 1, a_0 y_0)$ prior, where $y_0^* = y_0/\alpha$.

- ii) Suppose that y_1, \dots, y_n are i.i.d. exponential(u), where $u = -\theta$ and $\theta < 0$. Then

$$\begin{aligned}\pi(\theta) &\propto \exp \left\{ na_0 \left(\theta \bar{y}_0 - \log \left(\frac{1}{-\theta} \right) \right) \right\} \\ &= \theta^{na_0+1-1} \exp \{ na_0 \bar{y}_0 \theta \},\end{aligned}$$

which is equivalent to $u \sim \text{gamma}(na_0 + 1, na_0 \bar{y}_0)$.

We see that the conjugate priors for exponential family models are easy to derive and all have closed forms in the i.i.d. case.

When we turn to regression models (i.e., GLM's), the story changes. Recall that a GLM is obtained from an exponential family density by linking θ to a set of covariates. In particular, for the i^{th} observation $\theta_i = \theta(x'_i \beta) = \theta(\eta_i)$. With this regression connection, we can write the likelihood function of β (assuming $\phi = 1$) as

$$\begin{aligned} L(\beta) \propto p(y | \beta) &= \exp \left\{ \sum_{i=1}^n \{y_i \theta_i - b(\theta_i)\} \right\} \\ &\equiv \exp \left\{ \sum_{i=1}^n \{y_i \theta(x'_i \beta) - b(\theta(x'_i \beta))\} \right\} \end{aligned}$$

We would now like to discuss some possible choices of prior distributions for β .

Noninformative Priors for β

The uniform prior, i.e., $\pi(\beta) \propto 1$, is always an attractive noninformative prior for β since it yields a posterior with parameters that match frequentist point estimates. Thus with $\pi(\beta) \propto 1$, the posterior of β is $p(\beta | y) \propto L(\beta)$, and in this case the posterior mode of β equals the maximum likelihood estimate of β .

Moreover, by the Bayesian central limit theorem,

$$\beta | y \rightarrow N\left(\hat{\beta}, I^{-1}(\hat{\beta})\right),$$

as $n \rightarrow \infty$, where

$$I(\beta) = X' \Delta V \Delta X|_{\beta=\hat{\beta}}$$

for GLM's, and $\hat{\beta}$ is the MLE of β .

Thus $\pi(\beta) \propto 1$ has lots of desirable properties as a noninformative prior for β .

Another noninformative prior for β is Jeffreys prior, given by

$$\pi(\beta) \propto |I(\beta)|^{\frac{1}{2}}.$$

For GLM's, $I(\beta) = X'\Delta V \Delta X$, so

$$\pi(\beta) \propto |X'\Delta V \Delta X|^{\frac{1}{2}}.$$

Note that Δ and V depend on β . That is, $\Delta \equiv \Delta(\beta)$, $V \equiv V(\beta)$. We see that Jeffreys prior for GLM's is, in general, quite different from the uniform prior.

It can be shown that for the binomial model, Jeffreys prior for β is proper for any link function. For the normal, Poisson, and gamma regression models, Jeffreys prior for β is an improper prior for any link function.

Note that for the usual normal model,

$$Y \sim N_n(X\beta, \sigma^2 I),$$

$$V = I, \quad \phi = \sigma^2, \quad \Delta = I.$$

If σ^2 is assumed known, then

$$\pi(\beta) \propto |\sigma^{-2}(X'X)|^{\frac{1}{2}} = \text{constant},$$

so that Jeffreys prior is a uniform prior for β in this case. For the binomial, Poisson, and gamma regression models, Jeffreys prior for β is NOT a uniform prior.

An article characterizing Jeffreys prior for GLM's and its implications on the posterior can be found in Ibrahim and Laud (1991, *JASA*).

Remark 4.1: Propriety of the Posterior Distribution for GLM's

1. Theorem 4.4

- If $\pi(\beta) \propto 1$ and the MLE of β exists for the GLM, then $p(\beta | y)$ is proper.
2. With Jeffreys prior, i.e., $\pi(\beta) \propto |X'\Delta V\Delta X|^{1/2}$, it can be shown that $p(\beta | y)$ is proper under some very general conditions. Note that Jeffreys prior is improper for the Poisson, gamma, and normal models. See Ibrahim and Laud (1991, *JASA*) for more details.
 3. If $\pi(\beta)$ is proper, then of course $p(\beta | y)$ is **always** proper for any GLM.

Informative Priors for β

The most common type of informative prior for β in GLM's is a normal prior, i.e., $\beta \sim N_p(\mu_0, \Sigma_0)$. If a previous study exists with historical data $D_0 = (n_0, y_0, X_0)$, then we can take

$$\pi(\beta | a_0) \propto [L(\beta | D_0)]^{a_0} \quad (4.6)$$

$$= \exp \left\{ a_0 \sum_{i=1}^{n_0} \{y_{0i}\theta(x'_{0i}\beta) - b(\theta(x'_{0i}\beta))\} \right\},$$

where $y_0 = (y_{01}, \dots, y_{0n_0})'$ is the $n_0 \times 1$ vector of responses for the historical data, X_0 is an $n_0 \times p$ matrix of covariates with i^{th} row x'_{0i} , n_0 is the sample size for the historical data, and a_0 is a specified hyperparameter.

Note that even when a_0 is fixed, this prior does NOT have a closed form in β . This is a computationally challenging prior to work with, in general. We can approximate the prior in (4.6) by a normal distribution. That is, as $n_0 \rightarrow \infty$, one can show that

$$\pi(\beta | a_0) \rightarrow N_p(\tilde{\beta}, a_0^{-1}\tilde{\Sigma}) , \quad (4.7)$$

where $\tilde{\beta}$ is the MLE of β based on the historical data $D_0 = (n_0, y_0, X_0)$, and

$$\tilde{\Sigma} = (X_0' \Delta_0 V_0 \Delta_0 X_0)^{-1} \Big|_{\beta=\tilde{\beta}} ,$$

where X_0 is the $n_0 \times p$ covariate matrix for the historical data, and Δ_0 and V_0 are Δ and V using the covariates from the historical data. The approximation in (4.7) is computationally easier to work with than the prior in (4.6).

We can establish the Bayesian central limit theorem for a GLM with an arbitrary prior $\pi(\beta)$ for β .

Theorem 4.5: Bayesian CLT for GLM's

Suppose $L(\beta)$ is the likelihood function of any GLM based on n observations, and suppose $\pi(\beta)$ is the prior for β . Then as $n \rightarrow \infty$,

$$\beta \mid y \rightarrow N_p(\beta^*, \Sigma^*) ,$$

where β^* is the posterior mode of β obtained by solving

$$\frac{\partial}{\partial \beta} [\log p^*(\beta \mid y)] = 0 ,$$

and

$$\Sigma^* = \left(-\frac{\partial^2}{\partial \beta \partial \beta'} [\log p^*(\beta \mid y)] \Big|_{\beta=\beta^*} \right)^{-1} ,$$

where $p^*(\beta \mid y) = L(\beta) \pi(\beta)$ is the unnormalized posterior density of β .

Remark 4.2

Assuming $\phi = 1$, note that

$$-\frac{\partial^2}{\partial \beta \partial \beta'} \log p^*(\beta | y) = X'(\Delta V \Delta - \dot{\Delta} H)X - \frac{\partial^2 \log \pi(\beta)}{\partial \beta \partial \beta'}$$

a) If $\pi(\beta) \propto 1$, then $\frac{\partial^2 \log \pi(\beta)}{\partial \beta \partial \beta'} = 0$.

b) If $\beta \sim N_p(\mu_0, \Sigma_0)$, so that

$$\pi(\beta) \propto \exp \left\{ -\frac{1}{2}(\beta - \mu_0)\Sigma_0^{-1}(\beta - \mu_0) \right\},$$

then $-\frac{\partial^2 \log \pi(\beta)}{\partial \beta \partial \beta'} = \Sigma_0^{-1}$, so that in this case

$$-\frac{\partial^2 \log p^*(\beta | y)}{\partial \beta \partial \beta'} = X'(\Delta V \Delta - \dot{\Delta} H)X + \Sigma_0^{-1}.$$

Remark 4.3

- a) If $\pi(\beta) \propto 1$, then $\beta^* = \hat{\beta} = \text{MLE of } \beta$.
- b) If $\pi(\beta) = N_p(\mu_0, \Sigma_0)$, then

$$\frac{\partial}{\partial \beta} [\log p^*(\beta \mid y)] = 0$$

$$\begin{aligned} &\Leftrightarrow \frac{\partial}{\partial \beta} [\log[L(\beta) \pi(\beta)]] = 0 \\ &\Leftrightarrow \frac{\partial}{\partial \beta} \log L(\beta) + \frac{\partial}{\partial \beta} \log \pi(\beta) = 0 \\ &\Leftrightarrow \frac{\partial}{\partial \beta} \log L(\beta) - \Sigma_0^{-1}(\beta - \mu_0) = 0 \\ &\Leftrightarrow X' \Delta S - \Sigma_0^{-1}(\beta - \mu_0) = 0. \end{aligned}$$

Thus, the posterior mode of β is found by solving

$$X' \Delta S - \Sigma_0^{-1}(\beta - \mu_0) = 0 .$$

Bayesian computations for GLM's can be easily conducted within Stan using normal priors for β .

Example 4.7: Chapman Data

These data were taken from a heart study conducted on $n = 200$ men.

$$y = \begin{cases} 1 & \text{If coronary incident in previous 10 years} \\ 0 & \text{otherwise} \end{cases}$$

3 covariates were measured: $x_1 = \text{age}$, $x_2 = \text{cholesterol level}$, $x_3 = \text{weight}$.

Let z_1, z_2, z_3 be the centered and scaled covariates, and let β be the vector of covariate effects (with an intercept).

Suppose the prior for β is $\pi(\beta) \propto 1$.

We will sample from the posterior distributions of β using (i) the **rstan** package, (ii) the **rstanarm** package, and (iii) GENMOD in SAS. The package **rstan** allows for the most flexibility when specifying the model, however, it often requires more time to compile the model than the **rstanarm** package and GENMOD.

- i) For the `rstan` package, we write out the model details in a `.stan` file.

The file `chapman.stan` is below:

```
data {  
    // Define variables in data  
    // Number of observations  
    int<lower=0> n;  
    // Number of beta parameters  
    int<lower=0> p;  
  
    // Covariates  
    vector[n] z1;  
    vector[n] z2;  
    vector[n] z3;  
  
    // Binary outcome  
    int<lower=0, upper=1> y[n];  
}  
  
parameters {  
    // Define parameters to estimate  
    real beta[p];  
}  
  
model {  
    // Prior part of Bayesian inference  
    // Flat prior for betas (no need to specify if non-informative)  
  
    // Likelihood part of Bayesian inference  
    y ~ bernoulli_logit(beta[1] + beta[2]*z1 + beta[3]*z2 + beta[4]*z3);  
}
```

R code:

```
# Read in data and center/scale covariates
file.stan <- "chapman.stan"
chapman <- read.table("chapman.txt", header = TRUE) [,c(1,4,6,7)]
y <- Chapman[,4]
Z <- scale(Chapman[,-4])
colnames(Z) <- c("z1", "z2", "z3")

# Arrange data for Stan
dat1 <- list()
dat1$z1 <- Z[,1]
dat1$z2 <- Z[,2]
dat1$z3 <- Z[,3]
dat1$y <- y
dat1$n <- nrow(Z)
dat1$p <- ncol(Z) + 1

# Run Stan code
library(rstan)
stan.code <- readChar(file.stan, file.info(file.stan)$size)
start.time1 <- Sys.time()
stan.mod1 <- stan( model_code = stan.code, data = dat1, chains = 4,
                   iter = 3000, warmup = 500, seed = 779 )
end.time1 <- Sys.time()

# Computation time
> end.time1 - start.time1
Time difference of 1.047633 mins
```

```

# Frequentist analysis - MLEs and SEs
freq.mod <- glm( y ~ Z, family = binomial() )
> round( summary(freq.mod)$coefficients[,1:2], 4 )

            Estimate Std. Error
(Intercept) -2.2445    0.2738
Zz1          0.6175    0.2426
Zz2          0.4239    0.2334
Zz3          0.4373    0.2063

# rstan results - posterior mean of betas
> print( stan.mod1, pars = c("beta"), digits = 4 )
4 chains, each with iter=3000; warmup=500; thin=1;
post-warmup draws per chain=2500, total post-warmup draws=10000.

      mean se_mean     sd   2.5%   25%   50%   75% 97.5% n_eff Rhat
beta[1] -2.3147  0.0032 0.2809 -2.9063 -2.4988 -2.3029 -2.1172 -1.8032  7934 1.0008
beta[2]  0.6373  0.0027 0.2469  0.1587  0.4714  0.6351  0.8004  1.1285  8260 1.0001
beta[3]  0.4358  0.0025 0.2401 -0.0312  0.2736  0.4373  0.5945  0.9106  9264 1.0000
beta[4]  0.4458  0.0023 0.2103  0.0327  0.3063  0.4456  0.5867  0.8516  8118 1.0001

```

- ii) The `rstanarm` package estimates previously compiled regression models using the `rstan` package and thus does not use a `.stan` file.

R code:

```
# Set up data frame
dat2 <- data.frame( intercept = rep(1,nrow(Z)), Z, y )

# Compile Stan model - setting "prior = NULL" indicates uniform prior
library(rstanarm)
fmla <- y ~ z1 + z2 + z3
start.time2 <- Sys.time()
stan.mod2 <- stan_glm( fmla, data = dat2, chains = 4,
                       iter = 3000, warmup = 500, prior = NULL,
                       family = binomial, seed = 779 )
end.time2 <- Sys.time()

> end.time2 - start.time2
Time difference of 4.272959 secs

# rstanarm results - posterior mean of betas
> summary( stan.mod2, probs = c(0.025, 0.25, 0.5, 0.75, 0.975), digits = 4 )

Estimates:
      mean     sd    2.5%   25%   50%   75%  97.5%
(Intercept) -2.3193 0.2828 -2.9062 -2.5012 -2.3104 -2.1216 -1.8019
z1           0.6394 0.2507  0.1574  0.4710  0.6341  0.8089  1.1373
z2           0.4373 0.2407 -0.0387  0.2783  0.4370  0.5995  0.9066
z3           0.4491 0.2155  0.0288  0.3037  0.4457  0.5893  0.8789
```

- iii) We fit the model in SAS using the GENMOD procedure:

```
* Read in data and standardize;
data chapman;
    infile "chapman.txt" firstobs=2;
    input x1 x2 x3 x4 x5 x6 y;
run;

proc standard data=chapman mean=0 std=1 out=chapman2;
    var x1 x4 x6;
run;

data chapman3;
    set chapman2;
    z1 = x1;
    z2 = x4;
    z3 = x6;
    drop x1-x6;
run;

* Fit model using GENMOD;
proc genmod data=chapman3 descending;
    model y = z1 z2 z3 / dist=bin link=logit;
    bayes seed=779 coeffprior=uniform;
run;
```

The results for the model using GENMOD with 10,000 iterations and a burn-in of 2000 are included in the table below.

Parameter	Mean	Standard Deviation	Percentile			95% HPD Interval	
			25%	50%	75%		
Intercept	-2.3228	0.2829	-2.5025	-2.3117	-2.1280	-2.8928	-1.8015
z1	0.6380	0.2481	0.4670	0.6385	0.8090	0.1401	1.1069
z2	0.4384	0.2450	0.2784	0.4375	0.6032	-0.0486	0.9156
z3	0.4465	0.2123	0.3005	0.4466	0.5863	0.0507	0.8720

Comparing the output from all three models, we see that `rstan`, `rstanarm`, and GENMOD provide similar posterior results.

Conjugate Priors for GLM's

We can use the ideas from Theorem 4.2 on page 452 to construct a class of conjugated priors for GLM's (see Chen & Ibrahim, 2003, Statistica Sinica).

Suppose y_0 is a prior prediction for y in the current experiment, and let a_0 denote our confidence in this prior prediction. Let $\phi = 1$ for ease of exposition and consider a prior of the form

$$\pi(\beta|y_0, a_0, X) \propto \exp\{a_0(y_0'\theta(X\beta) - J'b(\theta(X\beta)))\}, \quad (4.8)$$

where $J = (1, \dots, 1)'$ is a $n \times 1$ vector of ones.

Theorem 4.6

For the class of GLM's, $\pi(\beta|y_0, a_0, X)$ in (4.8) is a conjugate prior for β . Let $\pi(\beta|y_0, a_0, X)$ be denoted by $D(a_0, y_0, X)$. Then a posteriori, $p(\beta|y, y_0, a_0, x)$ is $D\left(a_0 + 1, \frac{a_0 y_0 + y}{a_0 + 1}, X\right)$.

Remark 4.4

Under a canonical link, $\theta(X\beta) = X\beta$, so that (4.8) reduces to

$$\pi(\beta|y_0, a_0, X) \propto \exp\{a_0(y_0'X\beta - J'b(X\beta))\}.$$

Remark 4.5

Note that the prior depends on the current covariate matrix X , and hence is different, but related to the power prior based on $D_0 = (n_0, y_0, X_0)$.

Remark 4.6

Since the prior depends on X , (a_0, y_0) have a different interpretation than that of the power prior based on $D_0(n_0, y_0, X_0)$. In the conjugate prior, y_0 is a prior prediction for y , and hence not historical data. Also, a_0 is a precision parameter reflecting the degree of confidence in our prior prediction y_0 .

Proof of Theorem 4.6

$$\begin{aligned}
 p(\beta|y, y_0, a_0, X) &\propto L(\beta|y, X) \pi(\beta|y_0, a_0, X) \\
 &\propto \exp\{y'\theta(X\beta) - J'b(\theta(X\beta))\} \\
 &\quad \times \exp\{a_0(y_0'\theta(X\beta) - J'b(\theta(X\beta)))\} \\
 &= \exp\left\{(a_0 + 1) \left[\left(\frac{a_0 y_0' + y'}{a_0 + 1} \right) \theta(X\beta) - J'b(\theta(X\beta)) \right] \right\} \\
 &= D\left(a_0 + 1, \frac{a_0 y_0 + y}{a_0 + 1}, X\right)
 \end{aligned}$$

Remark 4.7

$p(\beta|y, y_0, a_0, X)$ does not have a closed form in general.

Example 4.8: Normal Model with Canonical Link

We have $y \sim N_n(X\beta, I)$ and thus

$$L(\beta|y, X) \propto \exp \left\{ -\frac{1}{2}(y - X\beta)'(y - X\beta) \right\}.$$

The conjugate prior is given by

$$\begin{aligned}\pi(\beta|y_0, a_0, X) &\propto \exp \left\{ a_0 \left[y_0'X\beta - \frac{(X\beta)'(X\beta)}{2} \right] \right\} \\ &\propto \exp \left\{ -\frac{a_0}{2}(\beta - \mu_0)'(X'X)(\beta - \mu_0) \right\}\end{aligned}$$

where $\mu_0 = (X'X)^{-1}X'y_0$.

Thus, in this case

$$D(a_0, y_0, X) = N_p(\mu_0, a_0^{-1}(X'X)^{-1})$$

where $\mu_0 = (X'X)^{-1}X'y_0$.

Remark 4.8

$b(\theta) = \theta^2/2$ for the normal model. Thus the posterior distribution is given by

$$\begin{aligned} p(\beta|y, y_0, a_0, X) &\propto \exp \left\{ -\frac{1}{2}(y - X\beta)'(y - X\beta) \right\} \\ &\quad \times \exp \left\{ -\frac{a_0}{2}(\beta - \mu_0)'(X'X)(\beta - \mu_0) \right\} \\ &\propto \exp \left\{ -\frac{1}{2}(\beta - \tilde{\beta})'(a_0X'X + X'X)(\beta - \tilde{\beta}) \right\} \end{aligned}$$

where

$$\begin{aligned} \tilde{\beta} &= \Lambda\mu_0 + (I - \Lambda)\hat{\beta}, \\ \Lambda &= (X'X + a_0X'X)^{-1}(a_0X'X) \\ &= (a_0 + 1)^{-1}(X'X)^{-1}a_0(X'X) = \frac{a_0}{a_0 + 1}I \\ I - \Lambda &= I - \frac{a_0}{a_0 + 1}I = \frac{1}{a_0 + 1}I, \\ \mu_0 &= (X'X)^{-1}X'y_0 \\ \hat{\beta} &= (X'X)^{-1}X'y. \end{aligned}$$

Thus

$$\begin{aligned}\tilde{\beta} &= \left(\frac{a_0}{a_0 + 1} \right) (X'X)^{-1} X'y_0 + \left(\frac{1}{a_0 + 1} \right) (X'X)^{-1} X'y \\ &= (X'X)^{-1} X' \left(\frac{a_0 y_0 + y}{a_0 + 1} \right).\end{aligned}$$

Thus

$$D \left(a_0 + 1, \frac{a_0 y_0 + y}{a_0 + 1}, X \right) = N_p(\tilde{\beta}, (a_0 + 1)^{-1}(X'X)^{-1})$$

where $\tilde{\beta} = (X'X)^{-1} \left[\frac{a_0 y_0 + y}{a_0 + 1} \right]$.

Remark 4.9

The derivation is also straightforward if $y \sim N_n(X\beta, \sigma^2 I)$ with σ^2 known.

Example 4.9: Logistic Regression

We have

$$L(\beta|y, X) \propto \exp \{y'X\beta - J'b(X\beta)\}$$

$$\propto \exp \left\{ y'X\beta - \sum_{i=1}^n \log(1 + \exp \{x_i'\beta\}) \right\}$$

$$b(X\beta) = \begin{pmatrix} \log(1 + \exp \{x_1'\beta\}) \\ \vdots \\ \log(1 + \exp \{x_n'\beta\}) \end{pmatrix}_{n \times 1}, \quad X\beta = \begin{pmatrix} x_1'\beta \\ \vdots \\ x_n'\beta \end{pmatrix}_{n \times 1}$$

$J = (1, \dots, 1)'$ is an $n \times 1$ vector of ones,

$$y'X\beta = \sum_{i=1}^n y_i x_i'\beta.$$

Now

$$\begin{aligned}
 \pi(\beta|y_0, a_0, X) &\propto \exp \{a_0(y_0'X\beta - J'b(X\beta))\} \\
 &\propto \exp \left\{ a_0 \left[y_0'X\beta - \sum_{i=1}^n \log(1 + \exp \{x_i'\beta\}) \right] \right\} \\
 &= D(a_0, y_0, X)
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(\beta|y, y_0, a_0, X) &\propto L(\beta|y, X)\pi(\beta|y_0, a_0, X) \\
 &\propto \exp \left\{ y'X\beta - \sum_{i=1}^n \log(1 + \exp \{x_i'\beta\}) \right\} \\
 &\times \exp \left\{ a_0 y_0'X\beta - a_0 \sum_{i=1}^n \log(1 + \exp \{x_i'\beta\}) \right\} \\
 &= \exp \left\{ (a_0 + 1) \left[\left(\frac{a_0 y_0' + y'}{a_0 + 1} \right) X\beta - \sum_{i=1}^n \log(1 + \exp \{x_i'\beta\}) \right] \right\} \\
 &= \exp \left\{ (a_0 + 1) \left[\left(\frac{a_0 y_0' + y'}{a_0 + 1} \right) X\beta - J'b(X\beta) \right] \right\} \\
 &= D \left(a_0 + 1, \left(\frac{a_0 y_0' + y'}{a_0 + 1} \right), X \right).
 \end{aligned}$$

Example 4.10: Probit Regression

$$\begin{aligned}
 L(\beta|y, X) &\propto \exp \left\{ \sum_{i=1}^n \left\{ y_i \log \left(\frac{\Phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right) - \log \left(\frac{1}{1 - \Phi(x'_i \beta)} \right) \right\} \right\} \\
 \pi(\beta|a_0, y_0, X) &\propto \exp \left\{ \sum_{i=1}^n a_0 y_{0i} \log \left(\frac{\Phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right) - a_0 \log \left(\frac{1}{1 - \Phi(x'_i \beta)} \right) \right\} \\
 p(\beta|y, a_0, y_0, X) &\propto \exp \left\{ (a_0 + 1) \sum_{i=1}^n \left(\frac{a_0 y_{0i} + y_i}{a_0 + 1} \right) \log \left(\frac{\Phi(x'_i \beta)}{1 - \Phi(x'_i \beta)} \right) - \log \left(\frac{1}{1 - \Phi(x'_i \beta)} \right) \right\} \\
 &\propto \exp \left\{ (a_0 + 1) \left[\left(\frac{a_0 y'_0 + y'}{a_0 + 1} \right) \theta(X\beta) - J'b(\theta(X\beta)) \right] \right\} \\
 \theta(X\beta) &= \begin{pmatrix} \log \left(\frac{\Phi(x'_1 \beta)}{1 - \Phi(x'_1 \beta)} \right) \\ \dots \\ \log \left(\frac{\Phi(x'_n \beta)}{1 - \Phi(x'_n \beta)} \right) \end{pmatrix}_{n \times 1}, \quad b(\theta(X\beta)) = \begin{pmatrix} \log \left(\frac{1}{1 - \Phi(x'_1 \beta)} \right) \\ \dots \\ \log \left(\frac{1}{1 - \Phi(x'_n \beta)} \right) \end{pmatrix}_{n \times 1}
 \end{aligned}$$

Remark 4.10

The power prior with $D_0 = (n_0, y_0, X_0)$ along with a_0 is not conjugate in the sense described above. The reason for this is that X_0 is different from X and hence this ruins the conjugacy.

The power prior for GLM's, based on $D_0 = (n_0, y_0, X_0)$, is given by

$$\pi(\beta|a_0, y_0, X_0) \propto \exp \{a_0 [y'_0 \theta(X_0 \beta) - b(\theta(X_0 \beta))]\}.$$

Now

$$L(\beta|y, X) \propto \exp \{y' \theta(X\beta) - b(\theta(X\beta))\}.$$

and

$$p(\beta|y, y_0, X, X_0) \propto \exp \{a_0 y'_0 \theta(X_0 \beta) + y' \theta(X\beta) - a_0 b(\theta(X_0 \beta)) - b(\theta(X\beta))\}.$$

We cannot combine $\theta(X_0 \beta)$ and $\theta(X\beta)$ as before. The only way we can combine them is to let $X_0 = X$ (which also implies $n_0 = n$).

Remark 4.11

The power prior, based on $D_0 = (n_0, y_0, X_0)$, with $X_0 = X$ is a conjugate prior.

Logit link (canonical link)

$$\begin{aligned} g(\mu) &= \log\left(\frac{\mu}{1-\mu}\right) = \eta, \quad \eta = X\beta \\ \frac{\exp\{\eta\}}{1 + \exp\{\eta\}} &= \mu, \quad \theta = \eta \end{aligned}$$

Probit link

$$\begin{aligned} g(\mu) &= \Phi^{-1}(\mu) = \eta \\ \mu &= \Phi(\eta) \\ \theta &= \log\left(\frac{\Phi(\eta)}{1 - \Phi(\eta)}\right) \end{aligned}$$

Complementary log-log link

$$\begin{aligned}g(\mu) &= \log(-\log(1-\mu)) = \eta \\ \mu &= 1 - \exp\{-\exp\{\eta\}\} \\ \theta &= \log\left(\frac{1 - \exp\{-\exp\{\eta\}\}}{\exp\{-\exp\{\eta\}\}}\right) \\ &= \log(\exp\{\exp\{\eta\}\}(1 - \exp\{-\exp\{\eta\}\})) \\ &= \log(\exp\{\exp\{\eta\}\} - 1)\end{aligned}$$

so $\theta = \log(\exp\{\exp\{\eta\}\} - 1)$.

The likelihood function is

$$p(y_i|\theta_i) = \exp\{y_i\theta_i - b(\theta_i)\}$$

$b(\theta_i) = \log(1 + \exp\{\theta_i\})$ for binary data so for complementary log-log link we have

$$\begin{aligned} p(y_i|\beta) &= \exp\{y_i \log(\exp\{\exp\{x'_i\beta\}\} - 1) - \log(1 + \exp\{\log(\exp\{\exp\{x'_i\beta\}\} - 1)\})\} \\ &= \exp\{y_i \log(\exp\{\exp\{x'_i\beta\}\} - 1) - \log(1 + \exp\{\exp\{x'_i\beta\}\} - 1)\} \\ &= \exp\{y_i \log(\exp\{\exp\{x'_i\beta\}\} - 1) - \log(\exp\{\exp\{x'_i\beta\}\})\} \\ &= \exp\{y_i \log(\exp\{\exp\{x'_i\beta\}\} - 1) - \exp\{x'_i\beta\}\} \end{aligned}$$

$$\begin{aligned}
 p(y|\beta) &= \prod_{i=1}^n \exp \{y_i \log (\exp \{\exp \{x'_i \beta\}\} - 1) - \exp \{x'_i \beta\}\} \\
 &= \exp \left\{ \sum_{i=1}^n y_i [\log (\exp \{\exp \{x'_i \beta\}\} - 1) - \exp \{x'_i \beta\}] \right\} \\
 &= \exp \{y' \theta(X\beta) - J'b(\theta(X\beta))\}
 \end{aligned}$$

where $y = (y_1, \dots, y_n)', J' = (1, \dots, 1)$,

$$\theta(X\beta) = \begin{pmatrix} \log (\exp \{\exp \{x'_1 \beta\}\} - 1) \\ \vdots \\ \log (\exp \{\exp \{x'_n \beta\}\} - 1) \end{pmatrix}_{n \times 1},$$

$$b(\theta(X\beta)) = \begin{pmatrix} \exp \{x'_1 \beta\} \\ \vdots \\ \exp \{x'_n \beta\} \end{pmatrix}_{n \times 1}.$$

Bayesian Inference and Computation for GLM's

As we now know, the posterior distribution of β for a GLM does not have a closed form. Thus we have to rely on Gibbs sampling or some other MCMC method to sample from the posterior distribution.

For GLM's, the full (complete) conditional distributions do not have a closed form either, and thus we need to rely on rejection algorithms within the Gibbs sampler to sample from

$$p(\beta_j | y, \beta_i, i \neq j, i = 1, \dots, p).$$

It turns out that the class of GLM's has a very attractive property called **log-concavity**. That is, the log-likelihood function of a GLM is a concave function.

This log-concavity property will make rejection sampling easier. Thus, in a GLM, if the prior is also log-concave (i.e., $\log \pi(\beta)$ is a concave function), then the log posterior of β is a concave function since

$$\log p(\beta | y) = \log L(\beta) + \log \pi(\beta),$$

and the sum of log-concave functions is a concave function.

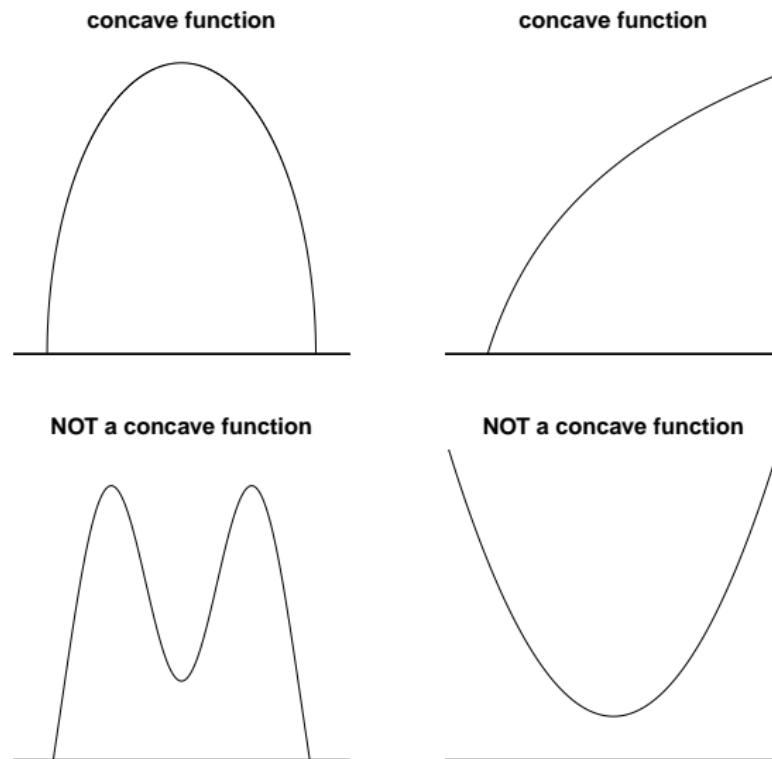


Figure 4.1:

Thus, if we choose $\pi(\beta)$ to be normal, then $\pi(\beta)$ is log-concave. A t prior for β is **NOT** log-concave in general.

Remark 4.12

A function $f(\theta)$ is log-concave if

$$\frac{\partial^2 \log f(\theta)}{\partial \theta^2} \leq 0 \quad .$$

For exponential family models, we have (assuming $\phi = 1$)

$$L(\theta) = \exp \{y\theta - b(\theta)\}$$

$$\log L(\theta) = y\theta - b(\theta)$$

$$\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = -b''(\theta) \leq 0 \quad ,$$

since $b''(\theta) = \text{Var}(y) \geq 0$.

Adaptive Rejection Sampling

This log-concavity property of GLM's has facilitated a new rejection algorithm for sampling from the complete conditionals. It is called the **adaptive rejection algorithm** and was developed by Gilks and Wild (1992, *Applied Statistics*).

Set $h(x) = \log[g(x)]$, where $g(x)$ is proportional to the density $f(x)$ from which we want to sample. In the Bayesian context, " $f(x)$ " = $p(\theta | x)$ is the posterior distribution. Let D be the domain of $f(x)$.

Let $\{x_i : i = 1, \dots, n\}$ denote points at which the function $h(x)$ has **already** been evaluated. Let $x_{(i)}$ denote the i^{th} lowest element of $\{x_i\}$. For example, $x_{(1)}$ is the lowest $x_{(i)}$, $x_{(2)}$ is the second lowest and so forth.

The lower hull, $h_l(x)$, is defined by the **chords** between $(x_{(i)}, h(x_{(i)}))$ and $(x_{(i+1)}, h(x_{(i+1)}))$.

The upper hull, $h_u(x)$, of $h(x)$ consists of the **tangents** at $\{x_i\}$. It is defined by the slope of the tangents $h'(x_i)$, the **abscissae** $\{z_{(i)}\}$ of the intersection between the tangents of $x_{(i)}$ and at $x_{(i+1)}$, and the values $h_u(z_{(i)})$.

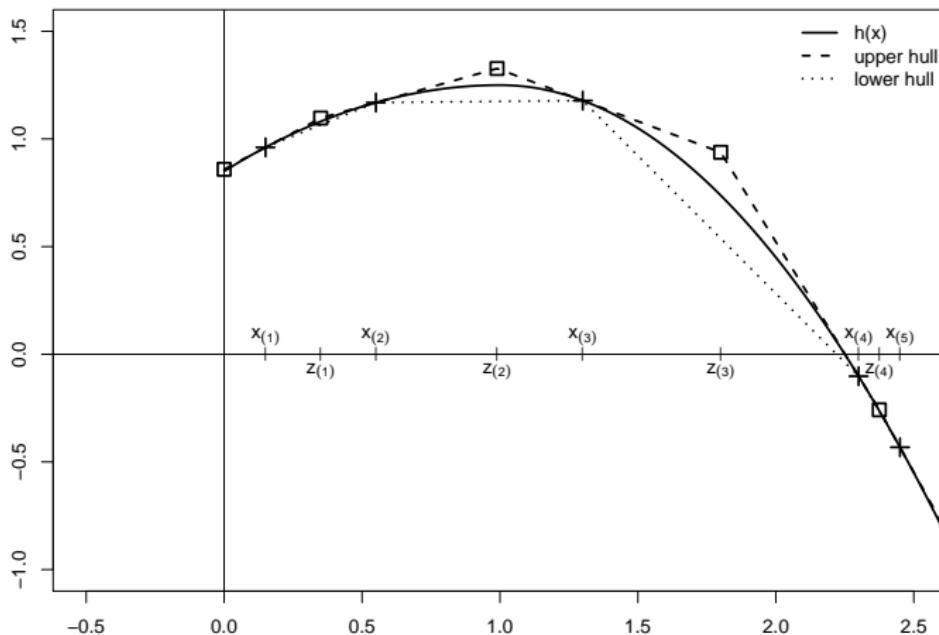


Figure 4.2: Concave log-density $h(x)$ (—), bounded on the left-hand side at $x = 0$, and by upper (—□—) and lower (··+··) hulls based on four abscissae ($z_{(1)}, z_{(2)}, z_{(3)}, z_{(4)}$).

Algorithm

Let A be a set of points $\{x_i, h(x_i), h'(x_i)\}$. The tangents at these points define the upper hull h_u , and the chords define the lower hull h_l (see Figure 4.2).

Step 1. Sample x from $s(x)$, the normalized exponential of the upper hull.

Step 2. Sample u from a uniform distribution over $(0, 1)$.

- ▶ If $u < \exp\{h_l(x) - h_u(x)\}$, then accept x without any function evaluation (squeezing);
- ▶ Else perform the rejection test.
 - ▶ Evaluate function h at x .
 - ▶ If $u < \exp\{h(x) - h_u(x)\}$, then accept x ;
 - ▶ Else reject x .
- ▶ If accepted, update lower and upper hull by adding $\{x, h(x), h'(x)\}$ to A , yielding closer upper and lower bounds for the subsequent rejection sampling.

Step 3. Repeat until the user-defined number of points have been sampled.

Starting Points

If there are m starting points, the algorithm will fail if one of the two following conditions is met:

- (a) $h'(x_{(1)}) \leq 0$, and there is no lower bound to the domain D of $f(x)$;
- (b) $h'(x_{(m)}) \geq 0$, and there is no upper bound to D .

In other words, if the domain D is not bounded, there must be starting points on both sides of the mode of $g(x)$; if D is unbounded on the left-hand side, the lowest starting point must be left of the mode; if D is unbounded on the right-hand side, the highest starting point must be to the right of the mode.

Updating and Sampling Formulae

To implement this algorithm we need to be able to calculate the abscissae and ordinates of the intersection points between the tangents

$$z_{(i)} = x_{(i)} + \frac{h(x_{(i)}) - h(x_{(i+1)}) + h'(x_{(i+1)})(x_{(i+1)} - x_{(i)})}{h'(x_{(i+1)}) - h'(x_{(i)})} ,$$

$$h_u(z_i) = h'(x_i)(z_i - x_i) + h(x_i) .$$

c_u is the normalizing factor of the exponentiated upper hull:

$$\begin{aligned} c_u &= \int \exp h_u(x) dx \\ &= \sum_{j=0}^{n-1} \frac{1}{h'(x_{(j+1)})} \{\exp h_u(z_{(j+1)}) - \exp h_u(z_{(j)})\} \end{aligned}$$

where $z_{(0)} = -\infty$ if there is no lower bound for x , $z_{(0)} = \text{xlb}$ if xlb is the lower bound of domain D , $z_{(n)} = +\infty$ if there is no upper bound for x and $z_{(n)} = \text{xub}$ if xub is the upper bound of domain D . Let us denote s_{cum} the cumulative distribution function of $s(x)$; then

$$s_{cum}(z_{(i)}) = \frac{1}{c_u} \sum_{j=0}^i \frac{1}{h'(x_{(j+1)})} \{\exp h_u(z_{(j+1)}) - \exp h_u(z_{(j)})\} .$$

To sample from $s(x)$, we sample first from a uniform distribution yielding u ; we then find the largest $z_{(i)}$ such that $s_{cum}(z_{(i)})$ is lower than u . The x -value sampled from $s(x)$ is then given by

$$x = z_{(i)} + \frac{1}{h'(x_{(i+1)})} \log \left[1 + \frac{h'(x_{(i+1)}) c_u \{u - s_{cum}(z_{(i)})\}}{\exp(h_u(z_{(i)}))} \right] .$$

Alternative Adaptive Rejection Sampling Algorithm

We now describe an alternative ARS algorithm developed by Gilks (1992). This alternative method differs from the Gilks and Wild (1992) method previously discussed.

Let $f(x)$ denote the density we wish to sample from. The domain D of f is an interval of the real line, densities are with respect to Lebesgue measure, and we define log-concavity of f as

$$\log[f(a)] - 2\log[f(b)] + \log[f(c)] < 0 \quad \forall a, b, c \in D \text{ such that } a < b < c.$$

This definition does not assume continuity in derivatives of f and includes, for example, linear and piecewise linear continuous functions.

Let $S_n = \{x_i; i = 0, \dots, n+1\}$ denote the *current* set of abscissae in ascending order, where x_0 and x_{n+1} are the possibly infinite lower and upper limits D . For $1 \leq i \leq j \leq n$, let $L_{ij}(x, S_n)$ denote the straight line through points $(x_i, \log[f(x_i)])$ and $(x_j, \log[f(x_j)])$, and for their (i, j) , let $L_{ij}(x, S_n)$ be undefined.

Define a piecewise linear function $h_n(x)$:

$$h_n(x) = \min[L_{i-1,i}(x, S_n), L_{i+1,i+2}(x, S_n)], \quad x_i \leq x \leq x_{i+1}, \quad (4.9)$$

where we notationally suppress the dependence of $h_n(x)$ on S_n . Here we establish the convention that if b is undefined, then $\min(a, b) = \min(b, a) = a$. As a consequence of the assumed log-concavity of $f(x)$, $h_n(x)$ is an envelope for $\log[f(x)]$, i.e. $h_n(x) \geq \log[f(x)]$ everywhere in D . This is illustrated for $n = 4$ in Figure 4.3. We can now perform rejection sampling with the sampling distribution given by

$$g_n(x) = \frac{1}{m_n} \exp \{h_n(x)\} \quad (4.10)$$

where

$$m_n = \int \exp \{h_n(x)\} dx.$$

Thus $g_n(x)$ is the normalized $h_n(x)$, i.e.,

$$\int g_n(x) dx = 1.$$

Note that $g_n(x)$ is a piecewise exponential distribution and can be sampled directly (Gilks and Wild, 1992).

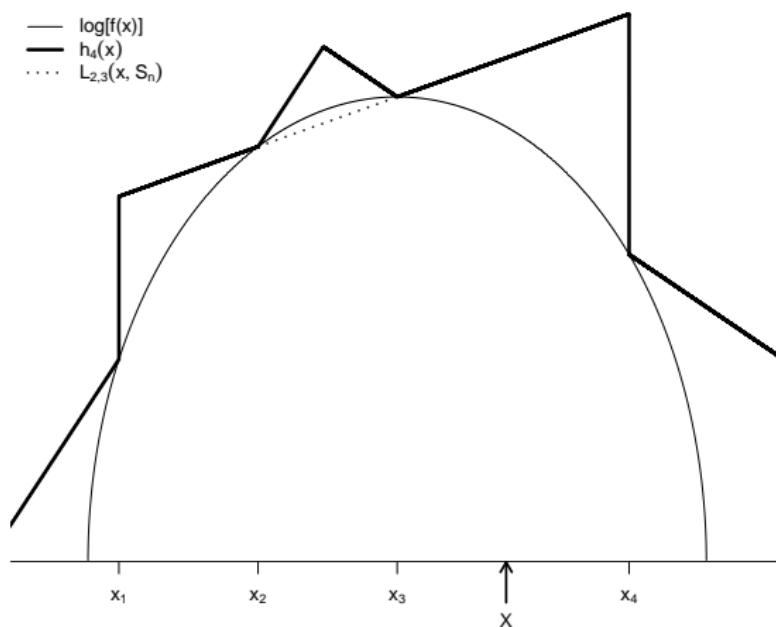


Figure 4.3: Adaptive rejection function $h_4(x)$ constructed from equation (4.9) for a log-concave function $f(x)$: X is sampled from $h_4(x)$.

The important feature of the sampling distribution $g_n(x)$ defined in equation (4.10) is that it can be updated each time that $f(X)$ is evaluated. We have then the following algorithm for ARS:

Step 0. Initialize n and S_n .

Step 1. Sample X from $g_n(x)$.

Step 2. Sample U from uniform(0, 1).

Step 3. Accept or reject X :

- ▶ If $U > f(X)/\exp\{h_n(X)\}$, then (*ARS rejection step*):
 - ▶ Set $S_{n+1} = S_n \cup \{X\}$,
 - ▶ Relabel points in S_{n+1} in ascending order,
 - ▶ Increment n and go back to step 1.
- ▶ Else set $X_A = X$ (*ARS acceptance step*).

Step 4. Return X_A .

At each iteration of ARS, the number of points of contact between $\log[f(x)]$ and $h_n(x)$ is increased by 1, thereby reducing m_n and decreasing the probability of rejection at step 3. This is illustrated in Figure 4.4. Further iterations of steps 1-4 will produce independent samples from f , while $h_n(x)$ is continually improving, making rejections increasingly less likely.

This method works straightforwardly if domain D is bounded on the left and right. If D is unbounded on the left, starting abscissae should be chosen so that the gradient of $L_{1,2}(x, S_n)$ is positive. Similarly, if D is unbounded on the right then the gradient of $L_{n-1,n}(x, S_n)$ should be negative.

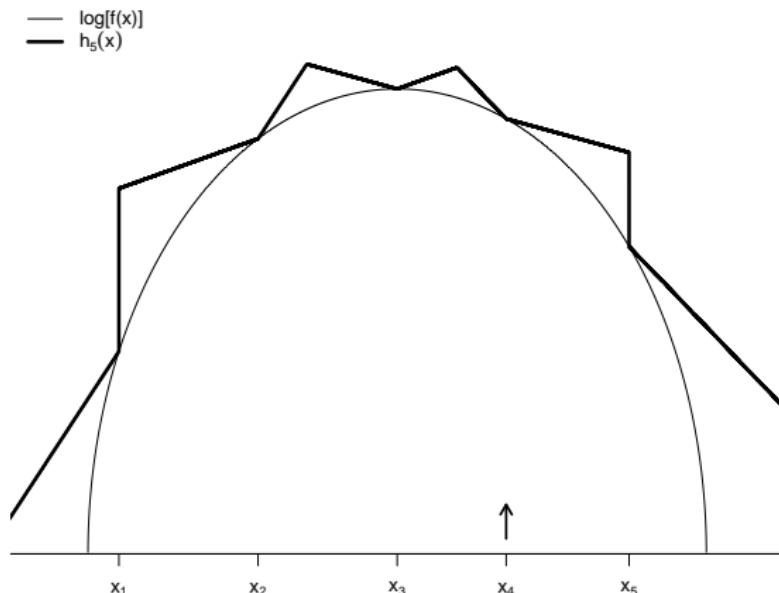


Figure 4.4: Updated current set S_5 and rejection function $h_5(x)$ after incorporating X in Figure 4.3.

The average number of iterations of ARS required to accept one point depends on the initial S_n and on f . Starting with $n = 3$, on average just two or three iterations are typically required, although with very poor starting values more may be necessary (Gilks, 1992). If f is a full conditional distribution, the envelope function $\exp \{h_n(x)\}$ from the previous iteration of the Gibbs sampler may be used to construct approximate 5%, 50%, and 95% centiles of f , for use as starting abscissae, although in many applications fixed starting abscissae will be adequate.

For densities $f(x)$ which are not log-concave, ARS cannot be used as $h_n(x)$ may not be an envelope for $\log \{f(x)\}$. To deal with non-log-concave densities we propose to append a Metropolis-Hastings algorithm step to ARS. We first briefly describe the Metropolis-Hastings algorithm.

Metropolis-Hastings Algorithm

The Metropolis algorithm (Metropolis et al., 1953) is, like the Gibbs sampler, an MCMC method. We describe the generalization of the algorithm given by Hastings (1970), which requires a *proposal distribution* $q(\cdot|z)$ from which samples X can be drawn for any z in D . The algorithm runs as follows:

Step 0. Set starting value x_0 ; set iteration counter $i = 0$.

Step 1. Sample X from $q(x|x_i)$.

Step 2. Sample U from uniform(0, 1).

Step 3. Accept or reject X :

- ▶ If $U > \min \left\{ 1, \frac{f(X)q(x_i|X)}{f(x_i)q(X|x_i)} \right\}$, then set $x_{i+1} = x_i$ (*rejection step*);
- ▶ Else set $x_{i+1} = X$ (*acceptance step*).

Step 4. Increment i and go back to step 1.

After suitably many iterations of this algorithm, the samples $\{x_i\}$ can be considered to be dependent samples from $f(x)$.

Tierney (1991) suggested the use of the Metropolis-Hastings algorithm within Gibbs sampling to sample from full conditional distributions. Indeed, this was the original form of the Metropolis algorithm (Metropolis et al., 1953). For this x_0 should be the value of x at the start of the current Gibbs iteration, and x_i will be the new value of x . Just one iteration of steps 1-4 suffices to preserve the stationary distribution of the Gibbs chain. However, this chain may be slower to converge through rejections at step 3.

Gilks, Best, Tan (1995, *Applied Statistics*) developed another algorithm to handle the cases when posterior densities are not log-concave . The algorithm involves introducing a Metropolis-Hastings step within the Gibbs sampler. They call the algorithm **Adaptive Rejection Metropolis Sampling** (ARMS). This alternative ARS method previously discussed is a special case of ARMS.

Adaptive Rejection Metropolis Sampling

We previously noted that ARS cannot be used to sample from non-log-concave distributions. To sample from such distributions, we could abandon rejection sampling in favor of the Metropolis-Hastings algorithm, applied to update one parameter (or one set of parameters) at a time. However, to avoid high probabilities of rejection (and hence slower convergence of chain) it may be helpful to adapt the proposal density q to the shape of the full conditional density f (Gelman, 1992). Since ARS provides a way of adapting a function f , we propose to use ARS to create a good proposal density for the Metropolis algorithm. We then append to ARS a single Metropolis-Hastings step, thus creating an ARMS within Gibbs chain.

However, unlike ARS, ARMS will not produce independent samples from f . ARMS is an adaptive generalization of the rejection sampling chain proposed by Tierney (1991).

Let $f(x)$ denote the density we wish to sample from.

Let (x, y) denote the complete set of variables being sampled by the Gibbs sampler. As before, x is the current variable to be sampled from its full conditional density (proportional to) $f(x)$, where we notationally suppress the conditioning on y . Let X_{cur} denote the current value of x at a given iteration of the Gibbs sampler. The aim then is to replace X_{cur} with a new value X_M from f .

For ARMS we construct a function $h_n(x)$ which is slightly more complex than in expression (4.9):

$$h_n(x) = \max[L_{i,i+1}(x, S_n), \min\{L_{i-1,i}(x, S_n), L_{i+1,i+2}(x, S_n)\}], \\ x_i \leq x \leq x_{i+1} \quad (4.11)$$

where, if b is undefined, $\min(a, b) = \min(b, a) = \max(a, b) = \max(b, a) = a$.

In general, $h_n(x)$ will not be an envelope of $\log[f(x)]$, as illustrated in Figure 4.5. The sampling density $g_n(x)$ is then given by (4.10) as before. Starting abscissae for ARMS must be independent of X_{cur} , as discussed below.

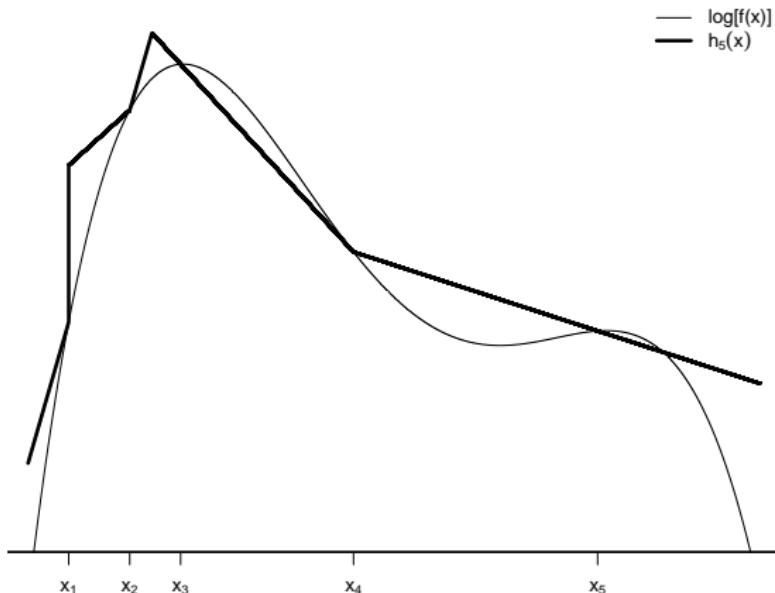


Figure 4.5: Adaptive function $h_5(x)$ for a non-log-concave function $f(x)$, constructed from equation (4.11)

The algorithm for ARMS then runs as follows:

Step 0. Initialize n and S_n independently of X_{cur} .

Step 1. Sample X from $g_n(x)$.

Step 2. Sample U from uniform(0, 1).

Step 3. Accept or reject X :

► If $U > f(X) / \exp\{h_n(X)\}$, then (ARS rejection step):

- Set $S_{n+1} = S_n \cup \{X\}$,
- Relabel points in S_{n+1} in ascending order,
- Increment n and go back to step 1.

► Else set $X_A = X$ (ARS acceptance step).

Step 4. Sample U from uniform(0, 1);

Step 5. Set X_M :

- If $U > \min \left[1, \frac{f(X_A) \min\{f(X_{cur}), \exp\{h_n(X_{cur})\}\}}{f(X_{cur}) \min\{f(X_A), \exp\{h_n(X_A)\}\}} \right]$, then set $X_M = X_{cur}$ (Metropolis-Hastings rejection step);
- Else set $X_M = X_A$ (Metropolis-Hastings acceptance step).

Step 6. Return X_M .

When f is log-concave then h_n in equation (4.11) reduces to expression (4.9) and is an envelope for $\log f$, so step 5 will always accept. Thus ARMS reduces to ARS for log-concave densities.

The proof that ARMS preserve the stationary distribution of the Gibbs sampler is an application of the auxiliary variables method (Besag and Green, 1993). The whole of the following argument conditions on y , so we shall not express this condition explicitly. Let N denote the value of n on reaching step 4. Then S_N determines the final state of the $h_n(x)$ -function. Conditionally on S_N and X_{cur} , X_A is a sample from

$$q(x|X_{cur}, S_N) \propto \min\{f(x), \exp\{h_n(x)\}\} \quad (4.12)$$

The starting abscissae for ARMS are drawn independently of X_{cur} . The right-hand side of expression (4.12) does not depend on X_{cur} . Thus we can consider $q(x|X_{cur}, S_N)$ as an independent Hastings-Metropolis proposal density, where S_N represents auxiliary variables. Let $P(X_M|X_{cur}, S_N)$ be the Markov transition function (for moving from X_{cur} to X_N) associated with the proposal density $q(x|X_{cur}, S_N)$ and the acceptance-rejection function in step 5 above.

Then it is straightforward to show that the detailed balance equation

$$f(X_{cur})P(X_M|X_{cur}, S_N) = f(X_M)P(X_{cur}|X_m, S_N) \quad (4.13)$$

holds for every S_N . By integrating equation (4.13) with respect to X_{cur} we see that X_M is independent of S_N and has density f . Thus X_M is a sample from the distribution of the full conditional for x , so ARMS within Gibbs sampling preserves the stationary distribution of the Gibbs chain.

This proof depends critically on the independence of S_N and X_{cur} . If starting abscissae were allowing to depend on X_{cur} we would need to replace $f(\cdot)$ in step 5 and equations (4.12) and (4.13) by the awkward density $f(\cdot|S_n)$, where

$$f(x|S_N) \propto f(x)pr(S_N|x = X_{cur}).$$

There is no need to iterate through steps 1-6 before updating y . However, the probability of moving away from X_{cur} can be increased through a fixed number of additional iterations of ARMS, setting $X_{cur} = X_M$, $n = N$ and $S_n = S_N$ in step 1 at the second and subsequent ARMS iteration. This is permissible because, as noted above, S_N and X_M are independent at the end of step 6, and hence X_{cur} and S_n will be independent at the start of the next ARMS iteration, as required by the theory. However, the stationary distribution of the ARMS within Gibbs chain would be affected if the decision to use extra iterations was allowed to depend in any way on previous rejections at step 5.

The probability of rejection at step 5 can be reduced through choosing good starting abscissae in S_n . An effective way of doing this is to base starting abscissae on the $\exp\{h_n(x)\}$ function from the previous Gibbs iteration, as described for ARS. This strategy for choosing starting abscissae is valid because $h_N(x)$ from the previous Gibbs iteration is independent of X_{cur} . This follows because X_{cur} is identically X_M from the previous Gibbs iteration, X_M and S_N are independent after each iteration of ARMS (as noted above) and S_N defines $h_N(x)$ completely.

This method works straightforwardly if domain D is bounded on the left and right. For D unbounded on the left or right, starting abscissae should be chosen as described for ARS.

Estimating Ratios of Normalizing Constants

A general approach to model selection and model comparison is that of estimating the ratio of two normalizing constants. There has been a lot of literature devoted to this problem, including Chen and Shao (1997, *Annals of Statistics*), Meng and Wong (1996, *Statistica Sinica*), Gelman and Meng (1994, 1996), Chib (1995, *JASA*), and Geyer (1994).

The basic problem is described as follows. Let $f_i(\theta)$, $i = 1, 2$, be two densities, each of which is known up to a normalizing constant. That is,

$$f_i(\theta) = \frac{f_i^*(\theta)}{c_i}, \quad \theta \in \Theta_i$$

where Θ_i is the parameter space for f_i , $i = 1, 2$.

The problem of estimating the ratio of two normalizing constants arises in the computation of Bayes factors and posterior model probabilities.

$$\begin{aligned} \text{BF} &= \frac{\int p(y | \theta_1) \pi(\theta_1) d\theta_1}{\int p(y | \theta_2) \pi(\theta_2) d\theta_2} \\ &= \frac{c_1}{c_2}, \end{aligned}$$

where c_1 is the normalizing constant of the posterior distribution of $[\theta_1 | y]$ and c_2 for $[\theta_2 | y]$. Here θ_1 and θ_2 denote parameter vectors from two different models.

Also, in the computation of posterior model probabilities, we have

$$\begin{aligned}
 p(m \mid y) &= \frac{p(y \mid m) p(m)}{\sum_{m \in \mathcal{M}} p(y \mid m) p(m)} \\
 &= \frac{\frac{p(y|m)}{p(y|m^*)} p(m)}{\sum_{m \in \mathcal{M}} \frac{p(y|m)}{p(y|m^*)} p(m)} \\
 &= \frac{\frac{c_m}{c_{m^*}} p(m)}{\sum_{m \in \mathcal{M}} \left(\frac{c_m}{c_{m^*}} \right) p(m)}.
 \end{aligned}$$

Let $r = \text{ratio of the two normalizing constants } c_1 \text{ and } c_2$. That is,

$$r = \frac{c_1}{c_2}.$$

We wish to estimate r , where c_i is the normalizing constant of density $f_i(\theta)$, i.e.,

$$c_i = \int_{\Theta_i} f_i^*(\theta) d\theta.$$

We will discuss several methods for estimating r .

Importance Sampling

This is the standard method, in which

$$\hat{r} = \frac{\frac{1}{n_1} \sum_{j=1}^{n_1} f_1^*(\theta_{1j}) / g_1(\theta_{1j})}{\frac{1}{n_2} \sum_{j=1}^{n_2} f_2^*(\theta_{2j}) / g_2(\theta_{2j})},$$

where g_1 and g_2 are two **importance sampling densities**.

Alternatively we can view r as an expectation

$$\begin{aligned} r &= \frac{c_1}{c_2} = E_2 \left[\frac{f_1^*(\theta)}{f_2^*(\theta)} \right] \\ &= \int \frac{f_1^*(\theta)}{f_2^*(\theta)} f_2(\theta) d\theta. \end{aligned} \tag{4.14}$$

If we can sample from $f_2(\theta)$, then r can be estimated by

$$\hat{r} = \frac{1}{n} \sum_{i=1}^n \frac{f_1^*(\theta_{2i})}{f_2^*(\theta_{2i})},$$

where $\theta_{21}, \theta_{22}, \dots, \theta_{2n}$ is a sample from $f_2(\theta)$. We note that if the two densities $f_1(\theta)$ and $f_2(\theta)$ have little overlap, the importance sampling procedure will work poorly.

Bridge Sampling

Bridge sampling uses a generalization of (4.14) and writes

$$r = \frac{c_1}{c_2} = \frac{E_2[f_1^*(\theta)\alpha(\theta)]}{E_1[f_2^*(\theta)\alpha(\theta)]},$$

where α is an arbitrary function defined on $\Theta_1 \cap \Theta_2$ such that

$$0 < \left| \int_{\Theta_1 \cap \Theta_2} \alpha(\theta) f_1^*(\theta) f_2^*(\theta) d\theta \right| < \infty.$$

Then, letting $\theta_{i1}, \theta_{i2}, \dots, \theta_{in_i}$ be a random sample from $f_i(\theta)$, $i = 1, 2$, we have

$$\hat{r}_\alpha = \frac{n_2^{-1} \sum_{i=1}^{n_2} f_1^*(\theta_{2i}) \alpha(\theta_{2i})}{n_1^{-1} \sum_{i=1}^{n_1} f_2^*(\theta_{1i}) \alpha(\theta_{1i})}.$$

Let $n = n_1 + n_2$ and $s_i = n_i/n$, $i = 1, 2$, and assume that $\lim_{n \rightarrow \infty} (s_i) > 0$, $i = 1, 2$. Meng and Wong (1996) showed that the optimal choice of α is given by

$$\alpha_{opt} = \frac{c}{s_1 f_1(\theta) + s_2 f_2(\theta)}, \quad \theta \in \Theta_1 \cap \Theta_2,$$

where $f_i(\theta) = f_i^*(\theta)/c_i$, $i = 1, 2$.

Note that α_{opt} minimizes the **relative mean square error**

$$\text{RE}^2(\hat{r}_\alpha) = \frac{E(\hat{r}_\alpha - r)^2}{r^2} .$$

Since c_1, c_2 are unknown, $\alpha_{opt}(\theta)$ is not directly usable.

As an alternative, Meng and Wong (1996) constructed the following iterative estimator,

$$\hat{r}_{opt}^{(t+1)} = \frac{\frac{1}{n_2} \sum_{i=1}^{n_2} f_1^*(\theta_{2i}) / (s_1 f_1^*(\theta_{2i}) + s_2 \hat{r}_{opt}^{(t)} f_2^*(\theta_{2i}))}{\frac{1}{n_1} \sum_{i=1}^{n_1} f_2^*(\theta_{1i}) / (s_1 f_1^*(\theta_{1i}) + s_2 \hat{r}_{opt}^{(t)} f_2^*(\theta_{1i}))} .$$

They showed that for each t ($t \geq 0$), $\hat{r}_{opt}^{(t+1)}$ provides a consistent estimator of r , and that the unique limit, \hat{r}_{opt} , achieves the asymptotic minimal relative mean-squared error.

The bridge sampling estimator will be unstable when $f_1^*(\theta)$ and $f_2^*(\theta)$ have little overlap. For such cases, the **path sampling** method of Gelman and Meng (1994) will substantially improve the simulation efficiency.

Path Sampling

Instead of trying to estimate $r = \frac{c_1}{c_2}$ directly, Gelman and Meng (1994) proposed the path sampling method to estimate the logarithm of r , that is,

$$\xi = -\log(r) = -\log\left(\frac{c_1}{c_2}\right).$$

Suppose we write

$$f_1(\theta) = \frac{f_1^*(\theta|\lambda_1)}{c_1(\lambda_1)},$$

$$f_2(\theta) = \frac{f_2^*(\theta|\lambda_2)}{c_2(\lambda_2)},$$

where λ_1 and λ_2 are hyperparameters that index $f_1(\theta)$ and $f_2(\theta)$. Suppose there exists a λ , such that when $\lambda_1 = \lambda_2 = \lambda$, we have

$$f_1(\theta) = \frac{f_1^*(\theta|\lambda_1)}{c_1(\lambda_1)} = \frac{f_2^*(\theta|\lambda_2)}{c_2(\lambda_2)} = f_2(\theta).$$

Denote this common density by $p(\theta|\lambda)$, and let $p^*(\theta|\lambda)$ denote the unnormalized $p(\theta|\lambda)$. That is,

$$p(\theta|\lambda) = \frac{p^*(\theta|\lambda)}{c(\lambda)} .$$

Let

$$U(\theta, \lambda) = \frac{\partial}{\partial \lambda} (\log[p^*(\theta|\lambda)]) ,$$

and let $\pi(\lambda)$ be a prior for λ on $[0, 1]$, and thus $0 \leq \lambda \leq 1$.

Gelman and Meng (1994) showed that

$$\xi = -\log \left(\frac{c_1}{c_2} \right) = E \left[\frac{U(\theta, \lambda)}{\pi(\lambda)} \right] ,$$

where the expectation is with respect to the joint density

$$f(\theta, \lambda) = f(\theta|\lambda)\pi(\lambda) .$$

Now let (θ_i, λ_i) , $i = 1, \dots, n$ be a random sample from $f(\theta, \lambda)$. Then

$$\hat{\xi} = \frac{1}{n} \sum_{i=1}^n \frac{U(\theta_i, \lambda_i)}{\pi(\lambda_i)} .$$

Here, $f(\theta, \lambda)$ serves as a **continuous path** to link $f_1(\theta|\lambda_1)$ and $f_2(\theta|\lambda_2)$. The Monte Carlo variance of $\hat{\xi}$ is

$$\text{Var}(\hat{\xi}) = \frac{1}{n} \left(\int_0^1 \frac{E[U^2(\theta, \lambda)]}{\pi(\lambda)} d\lambda - \hat{\xi}^2 \right) ,$$

where the expectation is taken with respect to $f(\theta|\lambda)$.

Ratio Importance Sampling (RIS)

Chen and Shao (1997, Annals of Statistics) propose a new method called **Ratio Importance Sampling** (RIS) to estimate $r = \frac{c_1}{c_2}$. Let $\Theta = \Theta_1 \cup \Theta_2$ denote the entire parameter space and let $g(\theta)$ be an arbitrary density over Θ . Then

$$r = \frac{c_1}{c_2} = \frac{E_g[f_1^*(\theta)/g(\theta)]}{E_g[f_2^*(\theta)/g(\theta)]}, \quad (4.15)$$

where $f_1^*(\theta)$ and $f_2^*(\theta)$ are the unnormalized densities of interest, and the expectation is taken with respect to $g(\theta)$.

Note that if $g(\theta) = \frac{f_2^*(\theta)}{c_2}$, then Chen and Shao's identity leads to the identity

$$r = \frac{c_1}{c_2} = E_{f_2} \left[\frac{f_1^*(\theta)}{f_2^*(\theta)} \right], \quad (4.16)$$

where the expectation is with respect to $f_2(\theta)$. Thus (4.16) can be estimated by

$$\hat{r} = \frac{1}{n} \sum_{i=1}^n \frac{f_1^*(\theta_{2i})}{f_2^*(\theta_{2i})},$$

where $\theta_{21}, \dots, \theta_{2n}$ is a random sample from $g(\theta) = f_2(\theta) = \frac{f_2^*(\theta)}{c_2}$.

Therefore, (4.15) is a generalization of (4.16). Given samples $\theta_1, \dots, \theta_n$ from $g(\theta)$, the RIS estimator for r is

$$\hat{r}_g = \frac{\sum f_1^*(\theta_i)/g(\theta_i)}{\sum f_2^*(\theta_i)/g(\theta_i)}.$$

It can be shown that \hat{r}_g is a consistent estimator of r . Chen and Shao (1997) derive the optimal $g(\theta)$, denoted $g_{opt}(\theta)$, which minimizes the expected relative mean squared error.

We now adapt the methods of Chen (1994) and Chen and Shao (1997) to variable subset selection. Let \mathcal{M} denote the model space, m is an arbitrary model in \mathcal{M} , and let m^* denote the **full model**. Let $\beta^{(m)}$ denote the $k_m \times 1$ regression coefficient vector for model m , and let $\beta \equiv \beta^{(m^*)}$ for ease of notation.

Write $\beta = (\beta^{(-m)}, \beta^{(m)})$, where $\beta^{(-m)}$ is the part of β **not** in $\beta^{(m)}$. Using the result of Chen and Shao (1997), we have

$$\frac{c_m}{c_{m^*}} = \frac{p(y|m)}{p(y|m^*)} = E_{\beta|y} \left[\frac{L(\beta^{(m)})\pi(\beta^{(m)})w(\beta^{(-m)}|\beta^{(m)})}{L(\beta)\pi(\beta)} \right],$$

where the expectation is taken with respect to the posterior density of $[\beta|y]$, i.e., the full model posterior density.

The weight function $w(\beta^{(-m)}|\beta^{(m)})$ is a **completely known** conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$. Chen (1994) shows that the **best choice** of $w(\beta^{(-m)}|\beta^{(m)})$ is

$$w(\beta^{(-m)}|\beta^{(m)}) = p(\beta^{(-m)}|\beta^{(m)}, y).$$

Since a closed form for $p(\beta^{(-m)}|\beta^{(m)}, y)$ is typically not available, we follow an empirical procedure provided by Chen (1994) to select $w(\beta^{(-m)}|\beta^{(m)})$.

Specifically, using the posterior sample $\{\beta_{(j)}, j = 1, \dots, N\}$, we construct the posterior mean and covariance matrix, denoted $(\tilde{\beta}, \tilde{\Sigma})$, and then we choose $w(\beta^{(-m)}|\beta^{(m)})$ to be the conditional density of the normal distribution $N_p(\tilde{\beta}, \tilde{\Sigma})$ for $\beta^{(-m)}|\beta^{(m)}$. Thus following the Monte Carlo method of Chen and Shao (1997) and using the posterior sample $\{\beta_{(j)}, j = 1, \dots, N\}$ from the full model, the posterior probability of model m can be estimated by

$$\hat{p}(m|y) = \frac{\frac{1}{N} \sum_{j=1}^N \left(\frac{L(\beta_{(j)}^{(m)}) \pi(\beta_{(j)}^{(m)}) w(\beta_{(j)}^{(-m)}|\beta_{(j)}^{(m)})}{L(\beta_{(j)}) \pi(\beta_{(j)})} \right) p(m)}{\sum_{k=1}^{2^p} \frac{1}{N} \sum_{j=1}^N \left(\frac{L(\beta_{(j)}^{(k)}) \pi(\beta_{(j)}^{(k)}) w(\beta_{(j)}^{(-k)}|\beta_{(j)}^{(k)})}{L(\beta_{(j)}) \pi(\beta_{(j)})} \right) p(k)}$$

for $m = 1, \dots, 2^p$, where β is $p \times 1$, and $\beta_{(j)} = (\beta_{(j)}^{(-m)}, \beta_{(j)}^{(m)})$.

The weight function $w(\beta^{(-m)} | \beta^{(m)})$ has the form

$$w(\beta^{(-m)} | \beta^{(m)}) = (2\pi)^{-\frac{(p-k_m)}{2}} |\tilde{\Sigma}_{11.2m}|^{-\frac{1}{2}} \\ \times \exp \left\{ -\frac{1}{2} (\beta^{(-m)} - \tilde{\mu}_{11.2m})' \Sigma_{11.2m}^{-1} (\beta^{(-m)} - \tilde{\mu}_{11.2m}) \right\},$$

where

$$\Sigma_{11.2m} = \tilde{\Sigma}_{11m} - \tilde{\Sigma}_{12m} \tilde{\Sigma}_{22m}^{-1} \tilde{\Sigma}'_{12m},$$

$\tilde{\Sigma}_{11m}$ is the covariance matrix from the marginal distribution of $\beta^{(-m)}$, $\tilde{\Sigma}_{12m}$ consists of the covariances between $\beta^{(-m)}$ and $\beta^{(m)}$, and $\tilde{\Sigma}_{22m}$ is the covariance matrix of the marginal distribution of $\beta^{(m)}$ with respect to the joint normal distribution $N_p(\tilde{\beta}, \tilde{\Sigma})$ for the vector $\beta = \beta^{(m*)}$.

Also, we note that

$$\tilde{\mu}_{11.2m} = \tilde{\mu}^{(-m)} + \tilde{\Sigma}_{12m} \tilde{\Sigma}_{22m}^{-1} (\beta^{(m)} - \tilde{\mu}^{(m)}),$$

where $\tilde{\mu}^{(-m)}$ is the mean of the marginal distribution of $\beta^{(-m)}$ implied by the $N_p(\tilde{\beta}, \tilde{\Sigma})$ distribution.

We can see why Chen's method is quite powerful for variable selection problems.

1. We only need samples from the full model posterior distribution.
2. The weight function w is not hard to choose.

The advantage of Chen's formula for $\hat{p}(m|y)$ is that samples **only** from the full model are needed to compute $\hat{p}(m|y)$ for all $m \in \mathcal{M}$.

In a general model selection problem, we have

$$\frac{c_m}{c_m^*} = \frac{\sum_{j=1}^N \frac{L(\beta_{(j)}|m)\pi(\beta_{(j)}|m)}{g(\beta_{(j)})}}{\sum_{j=1}^N \frac{L(\beta_{(j)}|m^*)\pi(\beta_{(j)}|m^*)}{g(\beta_{(j)})}} ,$$

where g can be taken to be a normal density and $\beta_{(j)}, j = 1, \dots, N$ are samples from g .

If such a g is hard to find and the models are not comparable, one could use Chen's modification of Chib's method. That, is c_m is estimated by

$$c_m = \left(\frac{1}{N} \sum_{j=1}^N \frac{w(\beta_{(j)}|m)}{L(\beta_{(j)}|m)\pi(\beta_{(j)}|m)} \right)^{-1},$$

where $\beta_{(j)}$, $j = 1, \dots, N$ are samples from the posterior distribution of $[\beta|y, m]$.

We note that this method is quite general and can be used for other types of models as well.

Example 4.11

We apply RIS to a logistic regression example with $n = 150$ observations and three covariates, x_{1i}, x_{2i}, x_{3i} , $i = 1, \dots, n$.

We simulate $y_i \sim \text{Bernoulli}(\mu_i)$ where $\mu_i = \frac{\exp\{1+2x_{1i}-2x_{3i}\}}{1+\exp\{1+2x_{1i}-2x_{3i}\}}$, $i = 1, \dots, n$, and $x_{ji} \sim N(0, \sigma_0^2)$, $j = 1, 2, 3$, where $\sigma_0 = 0.25$.

Let $\beta \sim N(0, 100^2)$, and assume uniform model priors.

R code to simulate data and fit model with Stan:

```
# Simulate data - three covariates and binary response
n <- 150
x1 <- numeric(n)
x2 <- numeric(n)
x3 <- numeric(n)
mu <- numeric(n)
y <- numeric(n)
p <- 4      # number of beta's

set.seed(779)
for(i in 1:n){
  x1[i] <- rnorm(1, 0, .25)
  x2[i] <- rnorm(1, 0, .25)
  x3[i] <- rnorm(1, 0, .25)
  mu[i] <- exp(1 + 2*x1[i] - 2*x3[i]) / (1 + exp(1 + 2*x1[i] - 2*x3[i]))
  y[i] <- rbinom(1, 1, mu[i])
}
z1 <- (x1 - mean(x1)) / sd(x1)
z2 <- (x2 - mean(x2)) / sd(x2)
z3 <- (x3 - mean(x3)) / sd(x3)
Z <- cbind(z1, z2, z3)
dat <- data.frame( intercept = rep(1,nrow(Z)), Z, y )
mod.mat <- as.matrix(dat[-5])

# Compile Stan model
library(rstanarm)
fmla <- y ~ z1 + z2 + z3
stan.full.mod <- stan_glm( fmla, data = dat, chains = 4,
                           iter = 3000, warmup = 500, prior = normal(0,100),
                           family = binomial, seed = 779 )
beta <- as.matrix(stan.full.mod)
```

```
# Model matrix indicating which covariates to include in each model
library(gtools)
ind.vec <- 0:1
all.mods <- permutations(n = 2, r = 3, v = ind.vec, repeats.allowed = TRUE)
# Note: last row of all.mods corresponds to full model
K <- nrow(all.mods) # number of models in model space
colnames(all.mods) <- c("x1", "x2", "x3")
rownames(all.mods) <- 1:K
> all.mods
   x1 x2 x3
1  0  0  0
2  0  0  1
3  0  1  0
4  0  1  1
5  1  0  0
6  1  0  1
7  1  1  0
8  1  1  1
```

R code to run model selection via RIS:

```

## Likelihood and Weights for Full Model

# Partition posterior mean and covariance of beta
beta.vals.full <- as.matrix(stan.full.mod)
beta.means.full <- colMeans(beta.vals.full)
beta.cov.full <- cov(beta.vals.full)
weights.full <- rep(1, dim(beta.vals.full)[1])      # weights for full model

# Calculate likelihood and prior values for qth sample of beta using full model
l.lik.full <- numeric(dim(beta.vals.full)[1])      # vector to store likelihood values
l.prior.beta.full <- numeric(dim(beta.vals.full)[1])      # vector to store prior values
for( j in 1:dim(beta.vals.full)[1] ){

  # Calculate components in likelihood
  beta.j <- beta.vals.full[j,]                      # Parameter values from sample j
  dens.y.full <- numeric( dim(mod.mat)[1] )
  for( i in 1:dim(mod.mat)[1] ){
    # Full model: success probability and density of y_il given beta^(j)
    lik.prob.full <- inv.logit( mod.mat[i,] %*% cbind(beta.j) )
    dens.y.full[i] <- lik.prob.full^y[i] * (1 - lik.prob.full)^(1 - y[i])
  }

  # Calculate likelihood for full models from sample j (log scale)
  l.lik.full[j] <- sum( log(dens.y.full) )

  # Prior density for beta from sample j (log scale)
  l.prior.beta.full[j] <- log( (2*pi)^(-p/2) ) + log( (det( 10000 * diag(p) ))^-0.5 ) +
    ( -0.5 * rbind(beta.j) %*% solve( 10000 * diag(p) ) %*% cbind(beta.j) )

  # Print iterations
  if( j %% 1000 == 0 ) print(j)
}

}

```

```

## Likelihood and Weights for All K - 1 Subset Models

# Store results for K - 1 models (last row of all.mods corresponds to full model)
weights.m <- matrix( 0, nrow = dim(beta.vals.full)[1], ncol = K - 1 )
summand.m <- matrix( 0, nrow = dim(beta.vals.full)[1], ncol = K - 1 )

# Weight function for m^th model
for(m in 1:(K - 1)){
  yes.betas <- which( c(1, all.mods[m,]) == TRUE )
  no.betas <- which( c(1, all.mods[m,]) == FALSE )
  k.m <- length(yes.betas)      # number of beta parameters in model m
  Sig.11.m <- as.matrix( beta.cov.full[no.betas,no.betas] )
  if(nrow(Sig.11.m) == 1){
    Sig.12.m <- rbind( beta.cov.full[no.betas,yes.betas] )
  }else{
    Sig.12.m <- cbind( beta.cov.full[no.betas,yes.betas] )
  }
  Sig.22.m <- as.matrix( beta.cov.full=yes.betas,yes.betas) )
  Sig.11.2.m <- Sig.11.m - Sig.12.m %*% solve(Sig.22.m) %*% t(Sig.12.m)

  for( j in 1:dim(beta.vals.full)[1] ){
    mu.11.2.m.j <- beta.means.full[no.betas] + Sig.12.m %*% solve(Sig.22.m) %*%
      ( cbind(beta.vals.full[j,yes.betas]) - cbind(beta.means.full=yes.betas) )
    weights.m[j,m] <- (2*pi)^(-(p - k.m)/2) * det(Sig.11.2.m)^-.5 *
      exp( -.5 * t(beta.vals.full[j,no.betas] - mu.11.2.m.j) %*% solve(Sig.11.2.m) %*%
        (beta.vals.full[j,no.betas] - mu.11.2.m.j) )
  }
}

```

```

# Estimator summands (w/o weights) from each sample j, j=1,...,N, for model m
for( j in 1:dim(beta.vals.full)[1] ){

  # Calculate components in likelihood
  beta.j <- beta.vals.full[j,]                      # Parameter values from sample j
  dens.y.m <- numeric( dim(mod.mat)[1] )
  for( i in 1:dim(mod.mat)[1] ){
    # Model m: success probability and density of y_il given beta^(j)
    lik.prob.m <- inv.logit( mod.mat[i,yes.betas] %*% cbind(beta.j[yes.betas]) )
    dens.y.m[i] <- lik.prob.m^y[i] * (1 - lik.prob.m)^(1 - y[i])
  }

  # Calculate likelihood for model m from sample j (log scale)
  l.lik.j.m <- sum( log(dens.y.m) )

  # Prior densities for beta from sample j (log scale)
  l.dens.beta.j.m <- log( (2*pi)^(-(k.m)/2) ) + log( (det( 10000 * diag(k.m) ))^-0.5 ) +
    ( -.5 * rbind(beta.j[yes.betas]) %*% solve( 10000 * diag(k.m) ) %*% cbind(beta.j[yes.betas]) )

  # Model m: jth summand in estimator (without weights)
  summand.m[j,m] <- exp( ( l.lik.j.m + l.dens.beta.j.m ) - ( l.lik.full[j] + l.prior.beta.full[j] ) )

  # Print iterations
  if( j %% 1000 == 0 ){
    print( paste(m, ":", j) )
  }
}

}

```

```
# Prior model probabilities
p.m <- rep(1/K, K - 1)      # uniform model priors
p.full <- 1/K

# Calculate posterior probabilities for all models
post.prob <- numeric(K)
denom.sum <- mean( 1 * weights.full ) * p.full
for(m in 1:(K - 1)){
  denom.sum <- denom.sum + mean( summand.m[,m] * weights.m[,m] ) * p.m[m]
}
for(m in 1:(K - 1)){
  post.prob[m] <- mean( summand.m[,m] * weights.m[,m] ) * p.m[m] / denom.sum
}
post.prob[K] <- 1 - sum(post.prob)
model.post.probs <- cbind( all.mods, PostProbability = round(post.prob, digits = 4) )
```

```
> model.post.probs
  x1 x2 x3 PostProbability
1  0  0  0      0.0236
2  0  0  1      0.0427
3  0  1  0      0.0002
4  0  1  1      0.0005
5  1  0  0      0.1386
6  1  0  1      0.7933
7  1  1  0      0.0010
8  1  1  1      0.0001
```

Stochastic Search Variable Selection

The computation of Bayes factors, HPD intervals, or posterior model probabilities will require MCMC techniques since the posterior distributions are not available in a closed form. It turns out that some novel MCMC algorithms can be developed for computing posterior model probabilities, in cases in which noninformative priors or informative priors are used. We now discuss some of these methods.

A popular method for computing posterior model probabilities using noninformative (but proper) priors was developed by George and McCulloch (1993, *JASA*), and George, McCulloch, and Tsay (1996). We now discuss these methods. The method discussed by George and McCulloch (1993) was written for the linear model but was extended to GLM's by George, McCulloch, and Tsay (1996).

George and McCulloch Model Selection Procedure

Suppose we have the following linear model

$$Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I).$$

They consider a prior for each β_i , $\beta = (\beta_1, \dots, \beta_p)'$ to be a mixture of two normal densities, and thus

$$\beta_i | \gamma_i \sim (1 - \gamma_i) N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2),$$

where γ_i is a binary random variable with

$$p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = p_i.$$

Notice in this specification that $\gamma_i = 1$ implies that the corresponding coefficient is important and should be included in the model, while $\gamma_i = 0$ implies that β_i is small enough that it can “safely” be estimated by 0.

Note that when $\gamma_i = 0$, $\beta_i \sim N(0, \tau_i^2)$ and when $\gamma_i = 1$, $\beta_i \sim N(0, c_i^2 \tau_i^2)$. The interpretation of this is as follows. Set τ_i ($\tau_i > 0$) small so that if $\gamma_i = 0$, then β_i would probably be so small that it could “safely” be estimated by 0. Second, if c_i ($c_i > 1$ always) is set large so that if $\gamma_i = 1$, then a non-zero estimate of β_i be probably included in the final model. Thus, the user must specify (τ_i, c_i) , $i = 1, \dots, p$.

Note here, that a priori, the β_i 's are **not** necessarily independent. Based on this interpretation, p_i may be thought of as the prior probability that β_i is **not** zero, or equivalently that X_i should be included in the model, where X_i denotes the i^{th} covariate.

The mixture prior for $\beta_i | \gamma_i$ can be written in vector form as

$$\beta | \gamma \sim N_p(0, D_\gamma R D_\gamma),$$

where $\gamma = (\gamma_1, \dots, \gamma_p)$, R is the prior correlation matrix and

$$D_\gamma = \text{diag}(a_1 \tau_1, \dots, a_p \tau_p),$$

where $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$. Thus D_γ determines the scaling of the prior covariance matrix.

The final piece of the model is to specify a prior for σ^2 . We take

$$\sigma^2 | \gamma \sim IG\left(\frac{v_\gamma}{2}, \frac{v_\gamma \lambda_\gamma}{2}\right).$$

The main reason for embedding the normal linear model in this hierarchical model is to obtain the **marginal posterior of γ** , i.e., $f(\gamma | y)$, where $f(\gamma | y) \propto f(y | \gamma) f(\gamma)$. The marginal posterior of γ contains the relevant information about variable selection.

The prior for γ can be taken as

$$f(\gamma) = \prod_{i=1}^p p_i^{\gamma_i} (1 - p_i)^{1-\gamma_i}.$$

To help guide the choice of c_i , it is useful to observe that the densities of $N(0, \tau_i^2)$ and $N(0, c_i^2 \tau_i^2)$ intersect at $\xi(c_i)\tau_i$ when $\xi(c_i) = \sqrt{\frac{2 \log(c_i)c_i^2}{c_i^2 - 1}}$.

This implies that the density of $N(0, c_i^2 \tau_i^2)$ will be larger than the density of $N(0, \tau_i^2)$ if and only if $|\beta_i| > \xi(c_i)\tau_i$. It may also be useful to observe that c_i is the ratio of the heights of $N(0, \tau_i^2)$ and $N(0, c_i^2 \tau_i^2)$ at 0. Thus c_i can be interpreted as the prior odds that X_i should be excluded when β_i is very close to 0. Choices for R include $R = I$ or $R \propto (X'X)^{-1}$.

Gibbs Sampling the Subsets

George and McCulloch (1993) propose a **stochastic search variable selection** (SSVS) to find the best possible subsets. The first part of SSVS entails specifying the hierarchical normal mixture model so that $f(\gamma | y)$ puts most weight on the more “promising” subsets of predictors. The second part of SSVS entails extracting this information. Rather than calculate all 2^p posterior probabilities in $f(\gamma | y)$, SSVS uses the Gibbs sampler to generate a sequence $\gamma^1, \dots, \gamma^n$, which in many cases converges rapidly in distribution to $\gamma \sim f(\gamma | y)$.

Such a sequence can be obtained quickly and efficiently, with far less effort than required to compute the entire posterior. The sequence $\gamma^1, \dots, \gamma^m$ will, with high probability in many cases, contain exactly the information relevant to variable selection. This is because those γ with highest probability will also appear most frequently and hence will be easiest to identify. These γ that appear infrequently or not at all are simply not of interest and can be disregarded. To see how SSVS works, let $\beta^0, (\sigma^2)^0, \gamma^0, \beta^1, \sigma^1, \gamma^1, \dots, \beta^j, \sigma^j, \gamma^j$ be a Gibbs sequence that is generated from the following scheme:

$$\begin{aligned}\beta^j &= f(\beta^j | y, (\sigma^2)^{j-1}, \gamma^{j-1}) \\ &= N_p((\sigma^{j-1})^{-2} A_{\gamma^{j-1}} X' X \hat{\beta}_{LS}, A_{\gamma^{j-1}})\end{aligned}\quad (4.17)$$

where $A_{\gamma^{j-1}} = ((\sigma^{j-1})^{-2} X' X + D_{\gamma^{j-1}}^{-1} R^{-1} D_{\gamma^{j-1}}^{-1})^{-1}$, and $\hat{\beta}_{LS}$ is the least squares estimate of β .

Next, the variance $(\sigma^2)^j$ is obtained by sampling

$$\begin{aligned} (\sigma^2)^j &\sim f((\sigma^2)^j \mid y, \beta^j, \gamma^{j-1}) \\ &= IG\left(\frac{n + v_{\gamma^{j-1}}}{2}, \frac{\|Y - X\beta^j\|^2 + v_{\gamma^{j-1}}\lambda_{\gamma^{j-1}}}{2}\right). \end{aligned} \quad (4.18)$$

Finally, the vector γ^j is obtained componentwise by sampling consecutively from the conditional distribution

$$\begin{aligned} \gamma_i^j &\sim f\left(\gamma_i^j \mid y, \beta^j, (\sigma^2)^j, \gamma_{(i)}^j\right) \\ &= f\left(\gamma_i^j \mid \beta^j, (\sigma^2)^j, \gamma_{(i)}^j\right), \end{aligned} \quad (4.19)$$

where $\gamma_{(i)}^j = (\gamma_1^j, \dots, \gamma_{i-1}^j, \gamma_{i+1}^{j-1} \dots \gamma_p^{j-1})$.

Notice that the distribution of γ_i^j does **not** depend on y . This simplification reduces computational requirements and allows for fast convergence of the sequence $\gamma^1, \dots, \gamma^m$.

Each γ_i^j is Bernoulli with

$$p\left(\gamma_i^j = 1 \mid \beta^j, (\sigma^2)^j, \gamma_{(i)}^j\right) = \frac{a}{a+b} ,$$

where

$$a = f(\beta^j \mid (\sigma^2)^j, \gamma_{(i)}^j, \gamma_i^j = 1) f((\sigma^2)^j \mid \gamma_{(i)}^j, \gamma_i^j = 1) f(\gamma_{(i)}^j, \gamma_i^j = 1)$$

and

$$b = f(\beta^j \mid (\sigma^2)^j, \gamma_{(i)}^j, \gamma_i^j = 0) f((\sigma^2)^j \mid \gamma_{(i)}^j, \gamma_i^j = 0) f(\gamma_{(i)}^j, \gamma_i^j = 0) .$$

By repeated successive sampling from (4.17), (4.18), and (4.19), the Gibbs sequence

$$\{\beta^j, (\sigma^2)^j, \gamma^j, j = 0, 1, \dots\}$$

is obtained.

As the length of this sequence is increased, the empirical distribution of the realized values of γ will converge to the actual posterior of $f(\gamma | y)$.

The sampled γ_i contain the relevant information about variable selection. The most frequently visited pattern of the γ_i corresponds to the best model. $\gamma_i = 1$ implies that the covariate should be included, while $\gamma_i = 0$ implies that the covariate should be excluded from the model.

The generalization of this method to GLM's is straightforward. The same priors are used, except we change $p(y | x, \beta)$ to a GLM. That is

$$p(y | x, \beta) \propto \exp \left[\sum_{i=1}^n \{y_i \theta(x'_i \beta) - b(\theta(x'_i \beta))\} \right].$$

The advantages of SSVS are that

- ▶ it can handle large problems,
- ▶ it is fast and efficient.

Disadvantages are that it does not use informative priors, nor allows incorporation of useful prior information.

Mitchell and Beauchamp (1988, *JASA*) also propose a method for doing Bayesian variable selection for the linear regression model using noninformative proper priors called **spike and slab** priors. Their method requires computation of all 2^p posterior model probabilities, and hence is not as popular as the method of George and McCulloch (1993).

Model Selection via MCMC & Related Methods

We want to study the computation of marginal likelihoods (or Bayes factors, posteriors model probabilities, ratios of normality constants) using MCMC methods.

In particular we want to characterize and compare single chain methods vs. methods that require an MCMC run for each model in the model space.

Methods requiring a single chain and no model enumeration

1. Reversible Jump – Green 1995; Brooks et.al. 2003, JRSSB; Richardson & Green, 1997 JRSS-B; Phillips & Smith, 1995; Markov Chain Monte Carlo Methods in Practice; Carlin & Louis 2006, chapter 6; Han & Carlin 2002, JASA
2. Stochastic Search Variable Selection (SSVS) – George and McCulloch 1993, JASA

Methods requiring a chain for each model and model enumeration

1. Importance sampling or Ratio Importance Sampling – Chen & Shao 1997, Annals of Statistics (not an MCMC method)
2. Marginal Likelihood – Chib, 1995; Chib & Jeliazkov, 2001; JASA
3. Newton-Raftery Method – Newton & Raftery, 1994 JRSS-B

Methods 2-3 are methods for computing the marginal likelihood alone, and method 1 compares ratios of normalizing constants.

Methods requiring one chain only and enumeration

1. RIS-IWMDE Method (ratios of normalizing constants) – Ibrahim, Chen, Maceachern 1999, CJS; Chen, Ibrahim, Yianmoutous, 1999, JRSS-B

We can also consider other methods, such as bridge sampling (Meng & Wang, 1996) (ratios of normalizing constants).

See the book by Chen, Shao, Ibrahim, (2000, Spring - Verlag) for more details on all of these methods.

Suppose we have two models, m_1 and m_2 , and the model space is denoted by \mathcal{M} . For now, we will focus on variable subset selection.

$$p(y|m_j) = \int p(y|\theta, m_j)\pi(\theta|m_j) d\theta, \quad j = 1, 2, \dots \quad (4.20)$$

and the Bayes factor for model m_1 over m_2 is given by

$$\text{BF} = \frac{p(y|m_1)}{p(y|m_2)}.$$

We discuss several methods to estimate $p(y|m_j)$, $j = 1, 2, \dots$

Newton - Raftery Method (also, Gelfand & Dey, 1994, JRSS-B)

$$\hat{p}(y|m) = \left[\frac{1}{N} \sum_{j=1}^N \frac{h(\theta^{(j)})}{p(y|\theta^{(j)}, m)\pi(\theta^{(j)}|m)} \right]^{-1},$$

where $\theta^{(j)}$ one samples from $p(\theta|y, m)$ obtained via MCMC, $h(\theta)$ is any proper density, i.e. $\int h(\theta)d\theta = 1$, $h(\theta) \geq 0$. Choices of h include multivariate normal or multivariate t densities, with mean equal to the posterior mean and covariance matrix equal to the posterior covariance matrix.

For example $h(\theta) = N(\theta^*, \Sigma^*)$ where

$$\theta^* = \frac{1}{N} \sum_{j=1}^N \theta^{(j)},$$

$$\Sigma^* = \frac{1}{N} \sum_{j=1}^N (\theta^{(j)} - \theta^*)(\theta^{(j)} - \theta^*)'.$$

Marginal Likelihood

If closed form full conditionals are available Chib (1995, JASA) presents an algorithm that avoids the specific of $h(\theta)$ discussed previously.

The method begins by writing

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

Only the denominator on the right side is unknown. Since the identity holds for any θ value, we require only a posterior density estimate at a single point, say θ' . Chib suggests picking θ' as the posterior mode. Thus, we have

$$\log \hat{p}(y|m) = \log p(y|\theta', m) + \log \pi(\theta'|m) - \log \hat{p}(\theta'|y, m). \quad (4.21)$$

How do we estimate $\hat{p}(\theta'|y, m)$?

Chib developed a **data augmentation** scheme (see Tanner and Wong, 1987, JASA) to estimate $\hat{p}(\theta'|y, m)$ by introducing latent variables. Chib's method is particularly useful for multivariate problems when the full conditional densities are completely known.

Suppose that $\theta = (\theta_1, \theta_2)$ (2 parameter blocks) where $p(\theta_1|y, \theta_2, m)$ and $p(\theta_2|y, \theta_1, m)$ are both available in closed form (Chib & Jeliazkov (2001) do not need a closed form for the full conditionals).

We write

$$p(\theta'_1, \theta'_2|y, m) = p(\theta'_2|y, \theta'_1, m)p(\theta'_1|y, m). \quad (4.22)$$

We observe that the first term on the right hand side is available explicitly at θ' , while the 2nd can be estimated as

$$\hat{p}(\theta'_1|y, m) = \frac{1}{N} \sum_{j=1}^N p(\theta'_1|y, \theta_2^{(j)}), \quad (4.23)$$

where $(\theta_1^{(j)}, \theta_2^{(j)}) = \theta^{(j)}$ are posterior samples from $p(\theta|y, m)$.

Thus, the marginal density estimate becomes

$$\begin{aligned} \log \hat{p}(y|m) &= \log p(y|\theta'_1, \theta'_2, m) + \log \pi(\theta'_1, \theta'_2|m) \\ &\quad - \log p(\theta'_2|y, \theta'_1, m) - \log \hat{p}(\theta'_1|y, m). \end{aligned}$$

Exponentiating produces the final marginal density estimate.

Next suppose there are three parameter blocks, $\theta = (\theta_1, \theta_2, \theta_3)$. The decomposition of the joint posterior density in (4.22) now becomes

$$p(\theta'_1, \theta'_2, \theta'_3 | y) = p(\theta'_3 | y, \theta'_1, \theta'_2) p(\theta'_2 | y, \theta'_1) p(\theta'_1 | y).$$

Again the first term is available explicitly, while the third term may be estimated as a mixture of $p(\theta'_1 | y, \theta_2^{(g)}, \theta_2^{(g)})$, $g = 1, \dots, G$ similar to (4.23) above. For the second term, we may write

$$p(\theta'_2 | y, \theta'_1) = \int p(\theta'_2 | y, \theta'_1, \theta_3) p(\theta_3 | y, \theta'_1) d\theta_3,$$

suggesting the estimator

$$\hat{p}(\theta'_2 | y, \theta'_1) = \frac{1}{G} \sum_{g=1}^G p(\theta'_2 | y, \theta'_1, \theta_3^{*(g)}),$$

where $\theta_3^{*(g)} \sim p(\theta_3 | y, \theta'_1)$. Such draws are *not* available from the original posterior sample, which instead contains $\theta_3^{(g)} \sim p(\theta_3 | y)$.

However, we may produce them simply by continuing the Gibbs sampler for an additional G iterations with only two full conditional distributions, namely

$$p(\theta_2|y, \theta'_1, \theta_3) \text{ and } p(\theta_3|y, \theta'_1, \theta_2)$$

Thus, while additional sampling is required, new computer code is not; we need only continue with a portion of the old code. The final marginal density estimate then arises from

$$\begin{aligned} \log \hat{p}(y) = & \log f(y|\theta'_1, \theta'_2, \theta'_3) + \log p(\theta'_1, \theta'_2, \theta'_3) \\ & - \log p(\theta'_3|y, \theta'_1, \theta'_2) - \log \hat{p}(\theta'_2|y, \theta'_1) - \log \hat{p}(\theta'_1|y). \end{aligned}$$

The extension to B parameter blocks requires a similar factoring of the joint posterior into B components, with $(B - 1)$ Gibbs sampling runs of G samples each to estimate the various factors. We note that clever partitioning of the parameter vector into only a few blocks (each still having a closed form full conditional) can increase computational accuracy and reduce programming and sampling time as well.

We can also estimate $\hat{p}(\theta' | y, m)$ using the **Importance-Weighted Marginal Density Estimation** (IWMDE) method of Chen (1994, JASA).

The IWMDE **does not** require completely known full conditional densities. Further, the IWMDE method can be used to estimate $p(y|m)$ directly.

Let $\theta_i, i = 1, \dots, n$ be a sample from $p(\theta|y)$. Such a sample can be obtained by MCMC methods, such as the Gibbs sampler or by the Metropolis-Hastings algorithm. Then IWMDE yields a consistent estimator for $p(y|m)$, given by

$$\tilde{p}(y|m) = \left(\frac{1}{n} \sum_{i=1}^n \frac{w(\theta_i)}{p(y|\theta_i, m)\pi(\theta_i|m)} \right)^{-1},$$

where $w(\theta)$ is a weighted completely known density function with the same support as that of the posterior distribution.

Chen (1994) discusses how to pick a good w . Chen's method for estimating $p(y|m)$ does *not* require obtaining a θ' . Chib's estimator works well if $\hat{p}(\theta'|y, m)$ is a good approximation and θ' is a good estimator for the mean (or mode). Chen's method works well if a good $w(\theta)$ is chosen.

Using Metropolis-Hastings output

Equations like (4.23) require us to know the normalizing constant for the full conditional distribution $p(\theta_1|y, \theta_2)$, thus precluding their use with full conditionals updated using Metropolis-Hastings (rather than Gibbs) steps.

To remedy this, Chib and Jeliazkov (2001) extend the approach, which takes a particular simple form in the case where the parameter vector θ can be updated in a single block. Let

$$\alpha(\theta, \theta^*|y) = \min \left\{ 1, \frac{p(\theta^*|y)q(\theta^*, \theta|y)}{p(\theta|y)q(\theta, \theta^*|y)} \right\},$$

the probability of accepting a Metropolis-Hastings candidate θ^* generated from a candidate density $q(\theta, \theta^*|y)$ (note that this density is allowed to depend on the data y). Chib and Jeliazkov (2001) then show

$$p(\theta'|y) = \frac{E_1\{\alpha(\theta, \theta'|y)q(\theta, \theta'|y)\}}{E_2\{\alpha(\theta', \theta|y)\}} \quad (4.24)$$

where E_1 is the expectation with respect to the posterior $p(\theta|y)$ and E_2 is the expectation with respect to the candidate density $q(\theta', \theta|y)$.

The numerator is then estimated by averaging the product in braces with respect to draws from the posterior, while the denominator is estimated by averaging the acceptance probability with respect to draws from $q(\theta', \theta|y)$, given the fixed value θ' .

Note that this calculation does not require knowledge of the normalizing constant for $p(\theta|y)$. Plugging the estimate of (4.24) into (4.21) completes the estimate of the marginal likelihood.

When there are two or more blocks, Chib and Jeliazkov (2001) illustrate an extended version of this algorithm using multiple MCMC runs, similar to the Chib (1995) approach for the Gibbs sampler previously outlined.

Bayes factors via sampling over the model space

The methods of the previous section all seek to estimate the marginal density $p(y)$ for each model, and subsequently calculate the Bayes factor via equation (4.20). They also operate on a posterior sample that has already been produced by some noniterative or Markov chain Monte Carlo method, though the methods of Chib (1995) and Chib and Jeliazkov (1999) will often require multiple runs of slightly different versions of the MCMC algorithm to produce the necessary output.

But for some complicated or high-dimensional model settings, such as spatial models using Markov random field priors (which involve large numbers of random effects parameters that cannot be analytically integrated out of the likelihood nor readily updated in blocks), these methods may be infeasible.

An alternative approach favored by many authors is to include the model indicator M as a parameter in the sampling algorithm itself. This of course complicates the initial sampling process, but has the important benefit of producing a stream of samples $\{M^{(g)}\}_{g=1}^G$ from $p(M|y)$, the marginal posterior distribution of the model indicator. Hence the ratio

$$\hat{p}(M = j|y) = \frac{\text{number of } M^{(g)} = j}{\text{total number of } M^{(g)}}, \quad j = 1, \dots, K \quad (4.25)$$

provides a simple estimate of each posterior model probability, which may then be used to compute the Bayes factor between any two of the models, say j and j' , via

$$\hat{\text{BF}}_{jj'} = \frac{\hat{p}(M = j|y)/\hat{p}(M = j'|y)}{p(M = j)/p(M = j')}, \quad (4.26)$$

which is the original formula used to define the Bayes factor.

Sampling over the model space alone

Most of the methods we will discuss involve algorithmic searches over both the model and the parameter-space simultaneously. This is of course a natural way to think about the problem, but like any other augmented sampler, such an approach risks a less well-identified parameter space, increased correlations, and hence slower convergence than a sampler operating on any one of the models alone.

Moreover, if our interest truly lies only in computing posterior model probabilities $p(M = j|y)$ or a Bayes factor, the parameter samples are not needed, relegating the θ_j to nuisance parameter status.

These thoughts motivate the creation of samplers that operate on the model space alone. This in turn requires us to integrate the θ_j out of the model before sampling begins. To obtain such marginalized expressions in closed form requires fairly specialized likelihood and prior settings, but several authors have made headway in this area from surprisingly broad classes of models.

For example, Madigan and York (1995) offer an algorithm for searching over a space of graphical models for discrete data, an approach they refer to as Markov chain Monte Carlo model composition, or $(MC)^3$.

Raftery, Madigan, and Hoeting (1997) work instead in the multiple regression setting with conjugate priors, again enabling a model-space-only search. They compare the $(MC)^3$ approach with the “Occam’s Window” method of Madigan and Raftery (1994).

Finally, Clyde, DeSimone, and Parmigiani (1996) use importance sampling (not MCMC) to search for the most promising models in a hierarchical normal linear model setting. They employ an orthogonalization of the design matrix which enables impressive gains in efficiency over the Metropolis-based $(MC)^3$ method.

Sampling over model and parameter space

Unfortunately, most model settings are too complicated to allow the entire parameter vector θ to be integrated out of the joint posterior in closed form, and thus require that any MCMC model search be over the model and parameter space jointly.

Carlin and Polson (1991) included M as a parameter in the Gibbs sampler. They computed Bayes factor and compared marginal posterior densities for a parameter of interest under changing specification of the model error densities and related prior densities. Their algorithm required that the models share the same parameterization, however, and so would not be appropriate for comparing two different mean structures (say, a linear and a quadratic).

George and McCulloch (1993) circumvented this problem for multiple regression models through stochastic search variable selection (SSVS), as previously discussed.

Unfortunately, in order to satisfy the Markov convergence requirement of the Gibbs sampler, a regressor can never completely “disappear” from the model, so the Bayes factors obtained necessarily depend on values of user-chosen tuning constants.

Reversible Jump MCMC

This method, originally due to Green (1995), is another strategy that samples over the model and parameter space, operating on the *union* space, $\mathcal{M} \times \cup_{j \in \mathcal{M}} \Theta_j$. It generates a Markov chain that can “jump” between models with parameter spaces of different dimensions, while retaining the aperiodicity, irreducibility, and detail balance conditions necessary for MCMC convergence.

A typical reversible jump algorithm proceeds as follows.

1. Let the current state of the Markov chain be (j, θ_j) , where θ_j is of dimension n_j .
2. Propose a new model j' with probability $h(j, j')$.
3. Generate u from a proposal density $q(u|\theta_j, j, j')$.

4. Set $(\theta_{j'}, u') = g_{j,j'}(\theta_j, u)$ where $g_{j,j'}$ is a deterministic function that is 1-1 and onto. This is a “dimension matching” function, specified so that $n_j + \dim(u) = n_{j'} + \dim(u')$.
5. Accept the proposed move (from j to j') with probability $\alpha_{j \rightarrow j'}$, which is the minimum of 1 and

$$\frac{f(y|\theta'_{j'}, M = j') p(\theta'_{j'}|M = j') \pi_{j'} h(j', j) q(u'|\theta_{j'}, j', j)}{f(y|\theta_j, M = j) p(\theta_j|M = j) \pi_j h(j, j') q(u|\theta_j, j, j')} \left| \frac{\partial g(\theta_j, u)}{\partial (\theta_j, u)} \right|$$

When $j' = j$, the move can be either a standard Metropolis-Hastings or Gibbs step. Posterior model probabilities and Bayes factor may be estimated from the output of this algorithm as described earlier.

The “dimension matching” aspect of this algorithm (step 4 above) is a bit obscure and merits further discussion. Suppose we are comparing two models, for which $\theta_1 \in \Re$ and $\theta_2 \in \Re^2$. If θ_1 is a subvector of θ_2 , then when moving from $j = 1$ to $j' = 2$, we might simply draw $u \sim q(u)$ and set

$$\theta'_2 = (\theta_1, u)$$

That is, the dimension matching g is the identity function, and so the Jacobian in step 5 is equal to 1. Thus if we set $h(1, 1) = h(1, 2) = h(2, 1) = h(2, 2) = 1/2$, we have

$$\alpha_{1 \rightarrow 2} = \min \left\{ 1, \frac{f(y|\theta'_2, M=2)p(\theta'_2|M=2)\pi_2}{f(y|\theta'_1, M=1)p(\theta_1|M=1)\pi_1q(u)} \right\},$$

with a corresponding expression for $\alpha_{2 \rightarrow 1}$.

In many cases, however, θ_1 will not naturally be thought of as a subvector of θ_2 . Green (1995) considers the case of *changepoint model* (see e.g. Table 5.7 and the associated exercises in Chapter 5 in Carlin and Louis), in which the choice is between a time series model having a single, constant mean level θ_1 , and one having two levels—say, $\theta_{2,1}$ before the changepoint and $\theta_{2,2}$ afterward. In this setting, when moving from Model 2 to 1, we would not likely want to use either $\theta_{2,1}$ or $\theta_{2,2}$ as the proposal value θ'_1 . A more plausible choice might be

$$\theta'_1 = \frac{\theta_{2,1} + \theta_{2,2}}{2}, \quad (4.27)$$

since the average of the pre- and post-changepoint levels should provide a competitive value for the single level in Model 1.

To ensure reversibility of this move, when going from Model 1 to 2 we might sample $u \sim q(u)$ and set

$$\theta'_{2,1} = \theta_1 - u \quad \text{and} \quad \theta'_{2,2} = \theta_1 + u$$

since this is a 1-1 and onto function corresponding to the deterministic down move (6.21).

Several variations or simplifications of reversible jump MCMC have been proposed for various model classes; see e.g. Richardson and Green (1997) in the context of mixture modeling, and Knorr-Held and Rasser (2000) for a spatial disease mapping application. Also, the “jump diffusion” approach of Phillips and Smith (1996) can be thought of as a variant on the reversible jump idea.

As with other Metropolis-Hastings algorithms, transformations to various parameters are often helpful in specifying proposal densities in reversible jump algorithms (say, taking the log of a variance parameter). It may also be helpful to apply reversible jump to a somewhat reduced model, where we analytically integrate certain parameters out of the model and use (lower-dimensional) proposal densities for the parameters that remain.

We will generally not have closed forms for the full conditional distributions of these “leftover” parameters, but this is not an issue since reversible jump does not require them. Here an example might be hierarchical normal random effects models with conjugate priors: the random effects (and perhaps even the fixed effects) may be integrated out, permitting the algorithm to sample only the model indicator and the few remaining (variance) parameters.

Summary and recommendations

Han and Carlin (2000) review and compare many of the methods described here in the context of two examples, the first a simple regression example, and the second a more challenging hierarchical longitudinal model (see Section 7.4 of Carlin and Louis). The methods described in this section that sample jointly over model and parameter space (such as reversible jump) often converge slowly, due to the difficulty in finding suitable pseudoprior (or proposal) densities.

Such methods are difficult to tune. The user will often need a rough idea of the posterior model probabilities $p(M = j|y)$ in order to set the prior model probabilities π_j in such a way that the sampler spends roughly equal time visiting each of the candidate models. Preliminary model-specific runs are also typically required to specify proposal (or pseudoprior) densities for each model.

By contrast, the marginal likelihood methods (Chib (1995), Chen (1994), Chen and Shao (1997), Newton and Raftery (1994)) appear relatively easy to program and tune. These methods do not require preliminary runs (only a point of high posterior density, θ'), and in the case of the Gibbs sampler, only a rearrangement of existing computer code. Estimating standard errors is more problematic (the authors' suggested approach involves a spectral density estimate and the delta method), but simply replicating the entire procedure a few times with different random number seeds generally provides an acceptable idea of the procedure's order of accuracy.

In their numerical illustrations, Han and Carlin (2000) found that the RJ method ran more quickly than the other space search methods, but the marginal likelihood methods seem to produce the highest degree of accuracy for roughly comparable runtimes. This is in keeping with the intuition that some gain in precision should accrue to MCMC methods that avoid a model space search.

As such, we recommend the marginal likelihood methods as relatively easy and safe approaches when choosing among a collection of standard (e.g., hierarchical linear) models.

We hasten to add, however, that the blocking required by these methods may preclude their use in some settings, such as spatial models using Markov random field priors. In such cases, reversible jump may offer the only feasible alternative for estimating a Bayes factor. The marginal likelihood methods would also seem impractical if the number of candidate models were very large (e.g., in variable selection problems with 2^p possible models, corresponding to each of p predictors being either included or excluded).

Model Checking

The Bayesian model checking tools are not as plentiful as for the linear model, but there are some basic guidelines to checking a model (GLM or not) that we state here. Gelman (Chapter 4) has an excellent chapter on model checking. Here are some basic notions in model checking.

One method is to compare your predictive distribution of future observations to data that have actually occurred. Construct your predictive distribution. Evaluate it at a few points and see if those look like your current data. Thus, we construct

$$p(z | y, x_f) ,$$

then evaluate $p(z | y, x_f)$ at several (z, x_f) and observe to see if these are consistent with your data. Another check involves drawing simulated values from the predictive distribution of replicated data and compare these samples to the observed data.

General Diagnostic Statistics

We can compute general discrepancy measures such as

$$T(y, \theta) = \sum (y_i - E(y_i | \theta))^2 / \text{Var}(y_i | \theta).$$

Then once $T(y, \theta)$ is computed, a Bayesian p-value can be computed as

$$\text{Bayes p-value} = P(T(y, \theta) \geq c | y).$$

Or perhaps a better way to compute the p-value is to take random draws from the predictive distribution $z | y$ and compute

$$\text{Bayes p-value} = P(T(z, \theta) \geq T(y, \theta) | y),$$

where the probability is taken over the posterior of θ . Yet another version of the Bayesian p-value would average over z , i.e.

$$\text{Bayes p-value} =$$

$$\int \int I(T(z, \theta) \geq T(y, \theta)) p(\theta | y) p(z | \theta) d\theta dz.$$

Bayesian p-values are tail areas under the posterior distribution, and thus are easy to interpret.

Sensitivity Analysis

Any type of model checking must also include sensitivity analyses with regards to changing prior distributions, prior parameters, and even the likelihood. One should check sensitivity with respect to

- ▶ posterior mean, mode, variance, quantiles
- ▶ HPD regions
- ▶ Bayes factors
- ▶ posterior model probabilities

In a sensitivity analysis, checking the likelihood itself is important. In some cases, the model is not good because of the **likelihood**. In these cases, we can consider **model expansion** (see Gelman et. al., p.177) by

- a) adding more parameters,
- b) a more general distribution (t instead of normal),
- c) a more general model that contains other models as special cases (as in the variable selection problem),
- d) considering multivariate response models instead of univariate response models.

CPO for GLM's

It can be shown that

$$\text{CPO}_i = \left\{ E_{\theta|y} \left(\frac{1}{p(y_i|\theta, x_i)} \right) \right\}^{-1}$$

so that a Monte Carlo estimate of CPO_i is

$$\hat{\text{CPO}}_i = \left(\frac{1}{N} \sum_{j=1}^N \frac{1}{p(y_i|\theta^{(j)}, x_i)} \right)^{-1}$$

where $\theta^{(1)}, \dots, \theta^{(N)}$ are samples from the posterior distribution of θ , i.e. $[\theta|y, x]$.

L measure for GLM's

$$\begin{aligned}
 L &= E_{z|y} \{(z - y)'(z - y)\} \\
 &= E_{\beta|y} \{E_{z|\beta} [(z - y)'(z - y)]\} \\
 &= E_{\beta|y} \left\{ \sum_{i=1}^n [\text{Var}(z_i|\beta) + (E(z_i|\beta) - y_i)^2] \right\} \\
 &= E_{\beta|y} \left\{ \sum_{i=1}^n [b''(\theta_i) + (b'(\theta_i) - y_i)^2] \right\}
 \end{aligned}$$

For logistic regression, the L measure is

$$L = E_{\beta|y} \left[\sum_{i=1}^n \frac{\exp \{x_i' \beta\}}{(1 + \exp \{x_i' \beta\})^2} + \left(\frac{\exp \{x_i' \beta\}}{1 + \exp \{x_i' \beta\}} - y_i \right)^2 \right].$$

Thus a Monte Carlo estimate of the L

$$\hat{L} = \frac{1}{N} \sum_{j=1}^N \left\{ \sum_{i=1}^n \frac{\exp \{x_i' \beta^{(j)}\}}{(1 + \exp \{x_i' \beta^{(j)}\})^2} + \left(\frac{\exp \{x_i' \beta^{(j)}\}}{1 + \exp \{x_i' \beta^{(j)}\}} - y_i \right)^2 \right\}$$

where $\beta^{(1)}, \dots, \beta^{(N)}$ are samples from $[\beta|x, y]$.

Chapter 5:

Bayesian Methods for Models with Longitudinal Data

Random Effects Models

Many longitudinal studies are designed to investigate changes over time in a characteristic that is measured repeatedly for each study participant. For example, in medical studies measurements such as blood pressure, cholesterol level, or lung volume may be taken on each individual at different time points and possibly changing experimental conditions. The most common type of model for repeated measurements is the linear random effects model of Laird and Ware (1982, *Biometrics*).

The model of Laird and Ware (1982) is described as follows: For a given individual i with n_i repeated measurements, the **random effects model** for outcome vector y_i is given by

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad i = 1, \dots, n,$$

where y_i is $n_i \times 1$, X_i is an $n_i \times p$ matrix of fixed covariates, β is a $p \times 1$ vector of regression coefficients, commonly referred to as **fixed effects**, Z_i is an $n_i \times q$ matrix of covariates for the $q \times 1$ vector of random effects b_i , and ϵ_i is an $n \times 1$ vector of random errors.

It is standard in implementations of this model to assume ϵ_i and b_i are independent, and both are normally distributed with

$$\epsilon_i \sim N_{n_i}(0, \sigma^2 I_{n_i})$$

and

$$b_i \sim N_q(0, D)$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix. Under these assumptions

$$y_i | \beta, b_i \sim N_{n_i}(X_i\beta + Z_i b_i, \sigma^2 I_{n_i}) .$$

Remark 5.1

In general, introducing random effects into a model creates a correlation structure between the components of y_i . Marginally,

$$\begin{aligned}y_i &= X_i\beta + Z_ib_i + \epsilon_i, \quad X_i \text{ is } n_i \times p, \quad Z_i \text{ is } n_i \times q \\y_i|b_i &\sim N(X_i\beta + Z_ib_i, \sigma^2 I_{n_i}), \quad i = 1, \dots, n.\end{aligned}$$

Given b_i , the components of y_i are independent.

Unconditionally, if $b_i \sim N_q(0, \sigma^2 D)$, then $y_i \sim N(X_i\beta, \sigma^2(I_{n_i} + Z_i D Z_i'))$. So now the components of y_i are dependent.

$$\begin{aligned}p(y_i) &= \int p(y_i|b_i)\pi(b_i) db_i \\&= N(X_i\beta, \sigma^2(I_{n_i} + Z_i D Z_i'))\end{aligned}$$

Frequentist Estimation of β and b_i

Laird and Ware (1982) discuss the estimation of (β, b_i, D) for this model and propose an EM algorithm for estimation.

Before going further into estimation, we need some motivation of the model. The random effects model can be viewed as a **two-stage model**. In this formulation, the distribution of the responses has the **same** form for each individual, but the parameters of that distribution vary over individuals. The distribution of these parameters (the b_i 's) are known as **random effects**, and constitutes the second stage of the model.

In a study of changes in lung volume during childhood, for instance, it may be reasonable to assume that the relationship between lung volume and the cube of height is linear in each child, but with linear regression parameters that vary among children. Such two stage models have desirable features. There is **no** requirement for balance in the data. They allow explicit modeling and analysis of between and within individual variation. **Repeated measures** and **growth curve** models are special cases of the random effects model. See Laird and Ware (1982, *Biometrics*) for further motivation of the random effects model.

Much of the theoretical work for random effects models was done by Harville (1977, *JASA*), Harville (1974, *Biometrika*), Harville (1976, *Annals of Statistics*). We can write the random effects model as a two-stage model by writing

► **Stage 1:**

$$y_i = X_i \beta + Z_i b_i + \epsilon_i , \quad (5.1)$$

where $\epsilon_i \sim N_{n_i}(0, R_i)$ and R_i can be taken as

$$R_i = \sigma^2 I_{n_i} , \quad i = 1, \dots, n .$$

► **Stage 2:**

$$b_i \sim N_q(0, D) , \quad (5.2)$$

where the b_i 's are i.i.d. and independent of the ϵ_i 's.

We call β the “fixed” effects and b_i the random effects. What distinguishes this model from the usual linear model is that we have a parameter vector for **each** individual.

Using the model (5.1) and (5.2), it can be shown that, **marginally**

$$y_i \mid \beta, R_i, D \sim N_{n_i}(X_i\beta, R_i + Z_i D Z_i') \quad (5.3)$$

where

$$p(y_i \mid \beta, R_i, D) = \int p(y_i \mid b_i, \beta, R_i, D) p(b_i \mid D) db_i ,$$

where $y_i \mid b_i, \beta, R_i, D \sim N_{n_i}(X_i\beta + Z_i b_i, R_i)$ and $b_i \mid D \sim N_q(0, D)$.

Inference for the random effects model usually proceeds from the **marginal model** (5.3), that is the model with the random effects integrated out.

Estimation with R_i and D known

Let $V_i = R_i + Z_i D Z_i' = \text{Var}(y_i | \beta, R_i, D)$ and let $W_i = V_i^{-1}$. In this case we can use weighted least squares to estimate β (and b_i). We are led to

$$\hat{\beta} = \left(\sum_{i=1}^n X_i' W_i X_i \right)^{-1} \left(\sum_{i=1}^n X_i' W_i y_i \right) .$$

Here $\hat{\beta}$ maximizes the marginal likelihood in (5.3) and is also the UMVUE of β .

We can use the estimate of β to estimate the b_i 's if we wish. We are led to

$$\hat{b}_i = D Z_i' W_i (y_i - X_i \hat{\beta}) .$$

Note that \hat{b}_i is **not** the MLE of b_i , but it has several nice properties as discussed by Harville (1976, *Annals of Statistics*).

Since $\hat{\beta}$ and \hat{b}_i are **linear** functions of y_i , their variances are easily derived as

$$\text{Var}(\hat{\beta}) = \left(\sum_{i=1}^n X_i' W_i X_i \right)^{-1},$$

$$\text{Var}(\hat{b}_i) = D Z_i' \left[W_i - W_i X_i \left(\sum_{i=1}^n X_i' W_i X_i \right)^{-1} X_i' W_i \right] Z_i D.$$

However, $\text{Var}(\hat{b}_i)$ understates the variance of $\hat{b}_i - b_i$, thus instead we use

$$\text{Var}(\hat{b}_i - b_i) = D - D Z_i' W_i Z_i D + D Z_i' W_i X_i \left(\sum_{i=1}^n X_i' W_i X_i \right)^{-1} X_i' W_i Z_i D.$$

We note that joint maximum likelihood estimates are very hard to obtain for the random effects model. There are no general efficient algorithms available.

Unknown R_i and D

Let $\theta = (R_1, \dots, R_n, D)$ be the vector of parameters for all of the variance components. The two most popular estimates are **maximum likelihood** and **restricted maximum likelihood** (REML). Thus we wish to jointly estimate (β, θ) .

To obtain ML estimates, Laird and Ware (1982) discuss the EM algorithm for the case $R_i = \sigma^2 I_{n_i}$. They view the problem as an incomplete data problem, where the b_i 's are viewed as **incomplete data**.

We note that the b_i 's are always **unobserved**. Let $\theta = (\sigma^2, D)$, and note that

$$\begin{aligned} & p(y|\beta, b)p(b|D) \\ & \propto \sigma^{-N} |D|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - X_i\beta - Z_i b_i)'(y_i - X_i\beta - Z_i b_i) - \frac{1}{2} \sum_{i=1}^n b_i' D b_i \right\} \\ & = \sigma^{-N} |D|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \epsilon_i' \epsilon_i - \frac{1}{2} \text{tr} \left(D \sum_{i=1}^n b_i b_i' \right) \right\} \end{aligned}$$

where $\epsilon_i = y_i - X_i\beta - Z_i b_i$.

Thus the complete sufficient statistic for σ^2 is $\sum_{i=1}^n \epsilon_i' \epsilon_i$, and the complete sufficient statistic for D is $\sum_{i=1}^n b_i b_i'$.

The EM algorithm proceeds as follows:

0) Start out with initial estimates $(\beta^{(0)}, \theta^{(0)})$.

1) **E-step**

At the $(s+1)^{st}$ EM iteration, compute

$$t_1^{(s+1)} = E \left[\sum_{i=1}^n \epsilon_i' \epsilon_i \mid y_i, \beta^{(s)}, \theta^{(s)} \right],$$

where $\epsilon_i = y_i - X_i \beta - Z_i b_i$. Also compute

$$t_2^{(s+1)} = E \left[\sum_{i=1}^n b_i b_i' \mid y_i, \theta^{(s)}, \beta^{(s)} \right].$$

2) **M-step**

Solve the equations

$$\sigma^2 {}^{(s+1)} = \frac{t_1^{(s+1)}}{N}, \quad N = \sum_{i=1}^n n_i$$

$$D^{(s+1)} = \frac{t_2^{(s+1)}}{N}.$$

3) Go back to step 1) with new estimates $\sigma^{2(s+1)}$ and $D^{(s+1)}$.

For the case of a general R_i , this EM algorithm will not work. In the general case, one would need a general optimization algorithm.

Note that

$$y_i \mid \beta, D, R_i \sim N_{n_i}(X_i\beta, V_i),$$

where $V_i = R_i + Z_i D Z'_i$.

The likelihood function based on **all** of the n observations is

$$\begin{aligned} L(\beta, \theta) &= \prod_{i=1}^n p(y_i \mid \beta, D, R_i) \\ &= \prod_{i=1}^n (2\pi)^{-\frac{n_i}{2}} |V_i|^{-\frac{1}{2}} \exp\{(y_i - X_i\beta)' V_i^{-1} (y_i - X_i\beta)\}, \end{aligned}$$

where $\theta = (R_1, \dots, R_n, D)$ and $V_i = R_i + Z_i D Z'_i$. Note that MLE's of θ and β do **not** have a closed form. For this general case, one could use Newton-Raphson to iteratively solve for (β, θ) . However, unless one specifies special structures for R_i and D , this estimation can be quite hard.

Typically, we take $R_i = \sigma^2 I_{n_i}$ and take D to be arbitrary. In this case, several algorithms have been proposed to obtain MLE's.

We note, however, that the MLE's for θ are **biased**. Several maximum likelihood types of algorithms have been proposed to produce **unbiased estimators** of variance components. One of the most popular methods is called **restricted maximum likelihood** (REML).

We now briefly describe the REML procedure. Let

$$N = \sum_{i=1}^n n_i, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{N \times 1}, \quad H = \begin{pmatrix} V_1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & V_n \end{pmatrix}_{N \times N},$$

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}_{N \times p}.$$

Then the marginal likelihood can be represented as

$$y \sim N_N (X\beta, H).$$

The REML estimator is defined as a maximum likelihood estimator based on a linearly transformed set of data $Y^* = AY$ such that the distribution of Y^* **does not** depend on β . One way to achieve this is by taking A to be

$$A_{N \times N} = I - X(X'X)^{-1}X'.$$

Then Y^* has a **singular** multivariate normal distribution with mean 0 regardless of β . To obtain a non-singular normal distribution for Y^* , we could use only $N - p$ rows of the matrix A . It turns out that the estimator of H does not depend on which row we use, nor any particular choice of A . Any full rank matrix with the property that $E(Y^*) = 0$ for all β will give the same answer.

Write $H \equiv H(\theta)$, where $\theta = (R_1, \dots, R_n, D)$. Define a matrix B such that $BB' = A$ and $B'B = I$, where B is $N \times N - p$. Finally let $Z = B'Y$.

Now for fixed θ , the maximum likelihood estimator for β is the generalized least squares estimator

$$\hat{\beta} = (X'H^{-1}X)^{-1}X'H^{-1}Y = GY ,$$

where $G = (X'H^{-1}X)^{-1}X'H^{-1}$.

Also, the respective densities of Y and $\hat{\beta}$ are given by

$$f(y | H, \beta) = (2\pi)^{-\frac{N}{2}} |H^{-1}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(y - X\beta)'H^{-1}(y - X\beta) \right\} ,$$

$$f(\hat{\beta}) = (2\pi)^{-\frac{p}{2}} |X'H^{-1}X|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\hat{\beta} - \beta)'(X'H^{-1}X)(\hat{\beta} - \beta) \right\} .$$

Note that

$$\begin{pmatrix} Z \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} B'_{N-P \times N} & Y \\ G_{p \times N} & Y \end{pmatrix} = CY, \quad C = \begin{pmatrix} B' \\ G \end{pmatrix}.$$

Since Z and $\hat{\beta}$ are linear functions of Y and Y is multivariate normal, to show Z and $\hat{\beta}$ are independent, it suffices to show $\text{Cov}(Z, \hat{\beta}) = 0$.

$$\begin{aligned} \text{Cov}(Z, \hat{\beta}) &= \text{Cov}(B'Y, (X'H^{-1}X)^{-1}X'H^{-1}Y) \\ &= B'\text{Cov}(Y) [H^{-1}X(X'H^{-1}X)^{-1}] \\ &= B'HH^{-1}X(X'H^{-1}X)^{-1} \\ &= B'X(X'H^{-1}X)^{-1}. \end{aligned}$$

Now since $B' \in C(X)^\perp$, $B'X = 0$ by definition. Thus Z and $\hat{\beta}$ are **independent** regardless of the value of β . Furthermore $E(Z) = 0$.

Now $\begin{pmatrix} Z \\ \hat{\beta} \end{pmatrix} = CY$, and thus $f(CY) \propto f(y)$ since CY is a linear transformation of Y .

Thus

$$\begin{aligned} f(Z, \hat{\beta} | H, \beta) &\propto f(y | H, \beta) \\ \Rightarrow f(Z | H, \beta) f(\hat{\beta}) &\propto f(y | H, \beta) \\ \Rightarrow f(Z | H, \beta) &\propto \frac{f(Y | H, \beta)}{f(\hat{\beta})}. \end{aligned}$$

Thus the density of Z , expressed in terms of y , is proportional to $\frac{f(y | H, \beta)}{f(\hat{\beta})}$.

That is

$$f(Z | H, \beta) \propto \frac{f(Y | H, \beta)}{f(\hat{\beta})}$$

where $Z = B'Y$.

An explicit form of this ratio is given by

$$\begin{aligned} \frac{f(y | H, \beta)}{f(\hat{\beta})} &= (2\pi)^{(N-p)/2} |H|^{-\frac{1}{2}} |X'H^{-1}X|^{-\frac{1}{2}} \\ &\times \exp \left\{ -\frac{1}{2} (y - X\hat{\beta})' H^{-1} (y - X\hat{\beta}) \right\}. \end{aligned} \quad (5.4)$$

We have made use of the fact that

$$(Y - X\beta)'H^{-1}(Y - X\beta) = (Y - X\hat{\beta})'H^{-1}(y - X\hat{\beta}) + (\hat{\beta} - \beta)'(X'H^{-1}X)(\hat{\beta} - \beta).$$

We note that the Jacobian of the transformation $Z = B'Y$ does not depend on any parameters in the model, and thus can be ignored. The practical implication of (5.4) is that the REML estimator $\tilde{\theta}$ maximizes the log-likelihood

$$\begin{aligned} L^*(\theta) &= -\frac{1}{2} \log |H| - \frac{1}{2} \log |X'H^{-1}X| \\ &\quad - \frac{1}{2}(y - X\hat{\beta})'H^{-1}(y - X\hat{\beta}), \end{aligned} \tag{5.5}$$

whereas the MLE of θ maximizes

$$L(\theta) = -\frac{1}{2} \log |H| - \frac{1}{2}(y - X\hat{\beta})' H^{-1}(y - X\hat{\beta}). \tag{5.6}$$

For a given H , we have

$$\hat{\beta} = \hat{\beta}(H) = (X'H^{-1}X)^{-1}X'H^{-1}y. \tag{5.7}$$

The REML estimator of H maximizes the log-likelihood in (5.5). We alternate between (5.5) and (5.7) to obtain the REML estimates of (β, θ) .

Bayesian Analysis of the Random Effects Model

Analysis proceeds with the marginal likelihood

$$y_i \mid \beta, R_i, D \sim N_{n_i}(X_i\beta, R_i + Z_i D Z_i') .$$

Letting $V_i = R_i + Z_i D Z_i'$, we have

$$\begin{aligned} p(y \mid \beta, R, D) &= \prod_{i=1}^n p(y_i \mid \beta, R_i, D) \\ &= \prod_{i=1}^n \left((2\pi)^{-\frac{n_i}{2}} |V_i|^{-\frac{1}{2}} \times \exp \left\{ -\frac{1}{2}(y_i - X_i\beta)' V_i^{-1} (y_i - X_i\beta) \right\} \right) \\ &= L(\beta, \theta), \end{aligned}$$

where $\theta = (R_1, \dots, R_n, D)$. Let us assume that $R_i = \sigma^2 I_{n_i}$, and take $b_i \sim N_q(0, \sigma^2 D)$ so that $V_i = \sigma^2(I_{n_i} + Z_i D Z_i')$.

Taking $b_i \sim N_q(0, \sigma^2 D)$ is convenient for computations as we will see.

Thus, we have

$$\begin{aligned} p(y | \beta, D) &= \prod_{i=1}^n (2\pi)^{-\frac{n_i}{2}} \sigma^{-n_i} |I_{n_i} + Z_i D Z_i'|^{-\frac{1}{2}} \\ &\times \exp \left\{ -\frac{1}{2\sigma^2} (y_i - X_i \beta)' (I_{n_i} + Z_i D Z_i')^{-1} (y_i - X_i \beta) \right\} \end{aligned}$$

Bayesian Analysis for D and σ^2 known

If D is known, then a Bayesian analysis proceeds in a straightforward way as if we had the usual linear model. Notice that

$$y | \beta, D \sim N_N(X\beta, \sigma^2 V) ,$$

where

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} , \quad V = \begin{pmatrix} V_1 & 0 \\ \ddots & \ddots & \ddots \\ 0 & \cdots & V_n \end{pmatrix} ,$$

$$V_i = I_{n_i} + Z_i D Z_i' , \quad N = \sum_{i=1}^n n_i .$$

Thus, if $\beta | \sigma^2 \sim N_p(\mu_0, \sigma^2 \Sigma_0)$, then

$$\begin{aligned} p(\beta | y, D, \sigma^2) &\propto p(y | \beta, D) \pi(\beta) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)' V^{-1} (y - X\beta) \right\} \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} [(\beta - \beta^*)' (X' V^{-1} X + \Sigma_0^{-1}) (\beta - \beta^*)] \right\}, \end{aligned}$$

where

$$\beta^* = (X' V^{-1} X + \Sigma_0^{-1})^{-1} \left((X' V^{-1} X) \hat{\beta} + \Sigma_0^{-1} \mu_0 \right),$$

and $\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} y$.

Thus

$$(\beta | y, D, \sigma^2) \sim N_p(\beta^*, \sigma^2 (X' V^{-1} X + \Sigma_0^{-1})^{-1}). \quad (5.8)$$

If one wishes to sample from the posterior distribution of β **without** deriving (5.8), we can use Gibbs sampling to do this.

Recall our model is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i .$$

To obtain samples from $[\beta | y, D, \sigma^2]$, we need to sample the full conditionals. Suppose $\beta | \sigma^2 \sim N_p(\mu_0, \sigma^2 \Sigma_0)$. Since $y | \beta, b, D, \sigma^2 \sim N_N(X\beta + Zb, \sigma^2 I)$, the full conditionals are straightforward to obtain.

Exercise 5.1

Derive the full conditionals. Note that

1. $\beta | b, D, \sigma^2, y \sim N_p(\cdot, \cdot)$
2. $b | \beta, D, \sigma^2, y \sim N_{nq}(\cdot, \cdot)$.

Cycling back and forth between 1) and 2), we obtain samples from the joint posterior distribution $p(\beta, b | D, \sigma^2, y)$. We then pick off the β samples from $\{(\beta_{(j)}, b_{(j)}), j = 1, \dots, N\}$, and these represent samples from the marginal posterior of β . That is, these are samples from

$$\begin{aligned} p(\beta | D, \sigma^2, y) &= \int p(\beta, b | D, \sigma^2, y) db \\ &= N_p(\beta^*, \sigma^2(X'V^{-1}X + \Sigma_0^{-1})^{-1}) . \end{aligned}$$

When D and σ^2 are known and $b_i \sim N_q(0, \sigma^2 D)$, the posterior distribution of β has a closed form. Specifically, it is a normal distribution.

When D is **unknown**, the models and priors become more complicated. Recall that D is a $q \times q$ positive definite matrix. For a Bayesian analysis, we need to specify a **prior distribution** for D . A common distribution to specify for covariance matrices is the **Wishart distribution** or the **inverse Wishart distribution**.

Definition 5.1

Suppose A is a $p \times p$ positive definite random matrix. The **Wishart distribution** with n degrees of freedom is characterized by

$$A \sim W_p(n, \Sigma)$$

if and only if

$$A = \sum_{i=1}^n y_i y_i' ,$$

where y_1, \dots, y_n are i.i.d. $N_p(0, \Sigma)$.

The density of A is given by

$$p(A) = k \frac{|A|^{\frac{1}{2}(n-p-1)} \exp\{-\frac{1}{2}\text{tr}(\Sigma^{-1}A)\}}{| \Sigma |^{n/2}}$$

$$\text{where } k = 2^{-\frac{np}{2}} \pi^{-\frac{p(p-1)}{4}} \left[\prod_{i=1}^p \Gamma\left(\frac{n+1-i}{2}\right) \right]^{-1}.$$

The density of A has $\frac{p(p+1)}{2}$ distinct random variables.

Example 5.1

Suppose x_1, \dots, x_N are i.i.d. $N_p(0, \Sigma)$.

Let

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i ,$$

$$S = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})' .$$

We have $\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})' \sim W_p(N-1, \Sigma)$ so that $(N-1)S \sim W_p(N-1, \Sigma)$.

Properties of Wishart Distribution

- 1) If $A \sim W_p(n, \Sigma)$, then $E(A) = n\Sigma$.
- 2) Diagonal elements of a Wishart matrix are chi-square random variables.
Suppose $A \sim W_p(n, \Sigma)$. Then

$$A_{ii} \sim \sigma_{ii}\chi_n^2 ,$$

where A_{ii} is the i^{th} diagonal element of A and σ_{ii} is the i^{th} diagonal element of Σ . Also, for any $\lambda \in R^p$,

$$\lambda' A \lambda \sim (\lambda' \Sigma \lambda) \chi_n^2 .$$

Similarly if L is a $p \times r$ matrix of constants,

$$L' A L \sim W_r(n, L' \Sigma L) .$$

- 3) The Wishart distribution has an additivity property. Suppose $A_i \sim W_p(n_i, \Sigma)$, $i = 1, \dots, k$, and the A_i 's are independent. Then

$$A_1 + \cdots + A_k \sim W_p \left(\sum_{i=1}^k n_i, \Sigma \right) .$$

Non-Central Wishart Distribution

Suppose $y_i \sim N_p(\mu_i, \Sigma)$, $i = 1, \dots, n$, and the y_i 's are independent. Let $W = \sum_{i=1}^n y_i y_i'$.

Then W has a **non-central** Wishart distribution, written

$$W \sim W_p(n, \Sigma, M'M) ,$$

where

$$M_{n \times p} = \begin{bmatrix} \mu_1' \\ \vdots \\ \mu_n' \end{bmatrix} .$$

In this case, we have $E(W) = n\Sigma + M'M$, and

$$W_{ii} \sim \sigma_{ii} \chi^2 \left(n, \frac{1}{\sigma_{ii}} \sum_{j=1}^n \mu_{ij}^2 \right) ,$$

where W_{ii} is the i^{th} diagonal element of W . If $L_{p \times r}$ is a matrix of constants, we have

$$L'WL \sim W_r(n, L'\Sigma L, L'M'ML) .$$

If $A \sim W_p(n, \Sigma)$, then

$$\psi_A(T) = |I - 2T\Sigma|^{-\frac{n}{2}} ,$$

where T is a $p \times p$ symmetric matrix.

Marginal Distributions

Suppose $A \sim W_p(n, \Sigma)$. Partition A as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where A_{11} is $r \times r$. Then

$$A_{11} \sim W_r(n, \Sigma_{11}).$$

Conditional Distributions

Suppose A has the partition above. Define

$$A_{11 \cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}.$$

Then

$$A_{11 \cdot 2} \sim W_r(n - p + r, \Sigma_{11 \cdot 2}).$$

The Inverse Wishart Distribution

Suppose $A \sim W_p(n, \Sigma)$. Then $B = A^{-1}$ is said to have an **inverse Wishart distribution**, denoted by

$$B \sim IW_p(n, \Sigma^{-1}) .$$

The density of B is given by

$$p(B) = \frac{|\Sigma^{-1}|^{\frac{n}{2}} |B|^{-\frac{1}{2}(n+p+1)} \exp\left\{-\frac{1}{2}\text{tr}(\Sigma^{-1}B^{-1})\right\}}{\pi^{\frac{p(p-1)}{4}} 2^{\frac{np}{2}} \prod_{i=1}^p \Gamma\left(\frac{n+1-i}{2}\right)} .$$

In the random effects model, when D is **unknown**, we typically specify a Wishart distribution for D^{-1} .

That is, we take

$$D^{-1} \sim W_q(\nu_0, C_0),$$

where (ν_0, C_0) are specified hyperparameters.

The joint posterior distribution of (β, D) does not have an analytic closed form when D is unknown. However, the **complete conditionals have closed forms**, which is attractive for Gibbs sampling.

Here, the complete conditionals that we need to obtain are (where $\tau = 1/\sigma^2$)

- 1) $[b | D, \beta, \tau, y]$ (Normal)
- 2) $[D^{-1} | b, \beta, \tau, y]$ (Wishart)
- 3) $[\beta | D, b, \tau, y]$ (Normal)
- 4) $[\tau | \beta, D, b, y]$ (gamma), $\tau = 1/\sigma^2$.

Exercise 5.2

Derive the complete conditionals in 1) - 4) above.

Thus, we see that Gibbs sampling for the random effects model is quite straightforward. All of the complete conditionals have a closed form, whereas the joint or marginal posterior distributions do not have an analytic closed form. Cycling through 1) - 4) yields samples from the joint posterior of $(b, D, \beta, \tau | y)$. Thus we obtain the samples

$$\{(b_{(j)}, D_{(j)}, \beta_{(j)}, \tau_{(j)}), j = 1, \dots, N\}.$$

Remark 5.2

The complete conditionals written on the previous page do not need to be in “vector form”. In scalar form, we write them as

- 1) $[b_{ij} | D, \beta, b_{i'j'}, \tau, y] \quad (i', j') \neq (i, j)$
- 2) $[D_{ij} | D_{i'j'}, \beta, b, \tau, y] \quad D_{ij} \neq D_{i'j'}$
- 3) $[\beta_j | \beta_k, D, \beta, b, \tau, y] \quad k \neq j$
- 4) $[\tau | \beta, D, b, y]$

Example 5.2: Normal Hierarchical Model with Rat Data

This example is taken from section 6 of Gelfand *et al.* (1990), and concerns 30 young rats whose weights were measured weekly for five weeks. Part of the data is shown below, where Y_{ij} is the weight of the i^{th} rat measured at age x_j .

	Weights Y_{ij} of rat i on day x_j				
	$x_j = 8$	15	22	29	36
Rat 1	151	199	246	283	320
Rat 2	145	199	249	293	354
.....					
Rat 30	153	200	244	286	324

A plot of the 30 growth curves suggests some evidence of downward curvature.

The model is essentially a random effects linear growth curve

$$\begin{aligned} Y_{ij} &\sim N(\alpha_i + \beta_i(x_j - \bar{x}), \tau_c^{-1}) \\ \alpha_i &\sim N(\alpha_c, \tau_\alpha^{-1}) \\ \beta_i &\sim N(\beta_c, \tau_\beta^{-1}) \end{aligned}$$

where $\bar{x} = 22$. We note the absence of a parameter representing correlation between α_i and β_i . For now, we standardize the x_j 's around their mean to reduce dependence between α_i and β_i in their likelihood.

We give $\alpha_c, \tau_\alpha, \beta_c, \tau_\beta, \tau_c$ the following independent “noninformative” priors:

$$\begin{aligned} \alpha_c &\sim N(0, 10000), \\ \beta_c &\sim N(0, 10000), \\ \tau_\alpha &\sim \text{gamma}(0.001, 0.001), \\ \tau_\beta &\sim \text{gamma}(0.001, 0.001), \\ \tau_c &\sim \text{gamma}(0.001, 0.001). \end{aligned}$$

Interest particularly focuses on the intercept at zero time (birth), denoted $\alpha_0 = \alpha_c - \beta_c \bar{x}$.

The first half of the file *rats.stan* is below:

```
data {  
    // Define variables in data  
    int N;          // Number of observations  
    int K;          // Number of time points  
    vector[K] x;    // Age (days)  
    matrix[N,K] Y; // Weight outcome  
    real xbar;      // Mean of ages when outcomes were measured  
}  
  
parameters {  
    // Define parameters  
    vector[N] alpha;  
    vector[N] beta;  
    real alphac;  
    real betac;  
    real<lower=0> tauc;  
    real<lower=0> taualpha;  
    real<lower=0> taubeta;  
}
```

The second half of the file *rats.stan* is below:

```
transformed parameters {
  real<lower=0> sigma2alpha = 1/taualpha;
  real<lower=0> sigma2beta = 1/taubeta;
  real<lower=0> sigma2c = 1/tauc;
  real alpha0 = alphac - betac * xbar;
}

model {
  // Random intercepts and slopes
  // Specify standard deviation (not variance or precision) in normal distribution
  alpha ~ normal(alphac, sqrt(sigma2alpha));
  beta ~ normal(betac, sqrt(sigma2beta));

  // Prior distributions
  alphac ~ normal(0, 100);
  betac ~ normal(0, 100);

  // Use rate parameterization for gamma distribution
  tauc ~ gamma(0.001, 0.001);
  taualpha ~ gamma(0.001, 0.001);
  taubeta ~ gamma(0.001, 0.001);

  // Likelihood part of Bayesian inference
  for(i in 1:N){
    for(j in 1:K){
      Y[i,j] ~ normal(alpha[i] + beta[i]*(x[j] - xbar), sqrt(sigma2c));
    }
  }
}
```

R code:

```
# Read in data and arrange it for Stan
file.stan <- "rats.stan"
Y <- read.table("rats.dat", header = FALSE)
x <- c(8, 15, 22, 29, 36)
colnames(Y) <- x
N <- nrow(Y)
K <- length(x)
dat <- list( Y = Y, x = x, N = N, K = K, xbar = mean(x) )

# Run Stan code
library(rstan)
stan.code <- readChar(file.stan, file.info(file.stan)$size)
stan.mod <- stan( model_code = stan.code, data = dat, chains = 4,
                  iter = 3000, warmup = 500, seed = 779 )

# Posterior mean of alpha_0 and beta_c
> print( stan.mod, pars = c("alpha0", "betac"), digits = 4 )

4 chains, each with iter=3000; warmup=500; thin=1;
post-warmup draws per chain=2500, total post-warmup draws=10000.

      mean    se_mean       sd     2.5%     25%     50%     75%   97.5% n_eff   Rhat
alpha0 106.3400  0.0350 3.5658 99.3115 103.9136 106.3255 108.7591 113.3409 10388 0.9997
betac   6.1888  0.0011 0.1077  5.9765  6.1186  6.1901  6.2584  6.4004  8970 0.9998
```

We estimate the average weight at birth to be about 106.

We now consider the problem of missing data. We delete the last observation of cases 6-10, the last two from 11-20, the last 3 from 21-25, and the last 4 from 26-30, as shown below.

```
151 199 246 283 320
145 199 249 293 354
.....
.....
153 NA NA NA NA
```

Stan does not support NA's in the outcome data frame Y , so we instead fit the model using JAGS with the R package `rjags`. JAGS is Just Another Gibbs Sampler, a program for analysis of Bayesian hierarchical models using MCMC simulation with similarities to BUGS. The user manual and additional examples for JAGS can be found at [sourceforge](#).

Not only does JAGS allow for missing values, but it also allows us to predict missing values. For this example, we are interested in the parameter estimates for α_0 and β_c and the predictions for the final four observations on rat 26.

We write out the model in the file *ratsJAGS.txt*, as shown below:

```
model{  
  # Likelihood part of Bayesian inference  
  for(i in 1:N){  
    for(j in 1:K){  
      Y[i,j] ~ dnorm(alpha[i] + beta[i]*(x[j] - xbar), tauc)  
    }  
  }  
  
  # Random intercepts and slopes  
  # Specify precision (not variance or standard deviation) in normal distribution  
  for(i in 1:N){  
    alpha[i] ~ dnorm(alphac, taualpha)  
    beta[i] ~ dnorm(betaac, taubeta)  
  }  
  
  # Prior distributions  
  alphac ~ dnorm(0, .0001)  
  betaac ~ dnorm(0, .0001)  
  # Use rate parameterization for gamma distribution  
  tauc ~ dgamma(0.001, 0.001)  
  taualpha ~ dgamma(0.001, 0.001)  
  taubeta ~ dgamma(0.001, 0.001)  
  
  alpha0 = alphac - betaac * xbar  
}
```

Prior to sampling from the model in R, we first must create a character vector with the names of all parameters and missing values which we want to sample (see the vector *jags.vars* below).

R code:

```
# Prepare input data for JAGS model
Y.miss <- read.table("ratsmiss.dat", header = FALSE)
data.jags <- list( Y = Y.miss, x = x, N = N, K = K, xbar = mean(x) )

# Set up model using JAGS ("rjags" package)
library(rjags)
seed.inits <- list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 779)
jags.mod <- jags.model( file = "ratsJAGS.txt", data = data.jags, inits = seed.inits,
                        n.adapt = 1000, n.chains = 1 )

# Specify names of parameters and missing values to sample
jags.vars <- numeric(6)
jags.vars[1] <- "alpha0"
jags.vars[2] <- "betac"
jags.vars[3] <- "Y[26,2]"
jags.vars[4] <- "Y[26,3]"
jags.vars[5] <- "Y[26,4]"
jags.vars[6] <- "Y[26,5]"
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )
```

```
# Summary of parameter estimates and missing value predictions  
> summary(jags.samps)
```

Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
Y[26,2]	204.304	9.3417	0.093417	0.127047
Y[26,3]	249.649	10.9700	0.109700	0.197734
Y[26,4]	294.837	13.5735	0.135735	0.281278
Y[26,5]	340.003	16.8820	0.168820	0.371789
alpha0	101.316	3.9136	0.039136	0.064327
betac	6.536	0.1932	0.001932	0.005155

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
Y[26,2]	186.57	198.43	204.46	210.247	221.86
Y[26,3]	229.09	242.72	249.75	256.711	270.01
Y[26,4]	270.30	286.68	295.08	303.343	319.26
Y[26,5]	310.41	330.21	340.26	350.114	370.01
alpha0	93.79	98.84	101.26	103.749	108.69
betac	6.25	6.44	6.54	6.637	6.83

The last four observed weights for rat 26 were 207, 257, 303 and 345, compared to our predictions of 204, 250, 295 and 340.

Prior and Posterior Distributions

Consider the random effects model

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i , \quad i = 1, \dots, n$$

$$b_i \sim N_q(0, \tau^{-1}D), \quad \epsilon_i \sim N_{n_i}(0, \tau^{-1}I_{n_i}),$$

where $\tau = 1/\sigma^2$. We can write this model as

$$Y = X\beta + Zb + \epsilon ,$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{N \times 1}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}, \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}_{N \times p},$$

$$Z = \begin{pmatrix} Z_1 & & 0 \\ & \ddots & \\ 0 & & Z_n \end{pmatrix}_{N \times nq}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}_{nq \times 1}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}_{N \times 1},$$

where $N = \sum_{i=1}^n n_i$.

Thus

$$\begin{aligned}\epsilon &\sim N_N(0, \tau^{-1} I_N), \\ b &\sim N_{nq}(0, \tau^{-1} (I_n \otimes D)).\end{aligned}$$

We assume a joint prior of the form

$$\pi(\beta, \tau, D^{-1}) \propto \pi(\beta|\tau) \pi(\tau) \pi(D^{-1})$$

or, if we also want to view the random effect as a “parameter”, we have the joint prior

$$\pi(\beta, b, \tau, D^{-1}) \propto \pi(\beta|\tau, D) \pi(\tau) \pi(b|\tau, D) \pi(D^{-1}).$$

Thus, we see that D^{-1} is independent of (β, τ) a priori.

A noninformative joint prior specification is

$$\pi(\beta, \tau, D^{-1}) \propto \tau^{-1} |D^{-1}|^{-\frac{(q+1)}{2}}. \quad (5.9)$$

Here a priori, β , τ , and D are all independent. It can be shown that the joint posterior distribution of (β, τ, D^{-1}) is proper with the improper prior given in (5.9).

The conjugate informative prior specification is

$$\beta | \tau \sim N_p(\mu_0, \tau^{-1} \Sigma_0), \quad (5.10)$$

$$\tau \sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right), \quad (5.11)$$

$$D^{-1} \sim W_q(\nu_0, C_0), \quad (5.12)$$

$$b | \tau, D \sim N_{nq}(0, \tau^{-1}(I_n \otimes D)). \quad (5.13)$$

We note that if $D^{-1} \sim W_q(\nu_0, C_0)$, then

$$\pi(D^{-1}) \propto |D^{-1}|^{\frac{1}{2}(\nu_0 - q - 1)} \exp\left\{-\frac{1}{2}\text{tr}(C_0^{-1}D^{-1})\right\}. \quad (5.14)$$

When $\nu_0 = 0$, $C_0^{-1} = 0$, we get the improper prior

$$\pi(D^{-1}) \propto |D^{-1}|^{-\frac{1}{2}(q+1)}.$$

The Wishart density in (5.14) is **proper** as long as $\nu_0 > q$ and C_0 is positive definite. Note that ν_0 is a scalar and C_0 is a $q \times q$ positive definite matrix.

The joint non-informative specification in (5.9) can be obtained from the informative specification in (5.10) - (5.13) by taking

$$\nu_0 = 0, \quad C_0^{-1} = 0, \quad \Sigma^{-1} = 0, \quad \delta_0 = 0, \quad \gamma_0 = 0.$$

In the Bayesian context $b \sim N_{nq}(0, \tau^{-1}(I_n \otimes D))$ is just viewed as another parameter. We can write the joint posterior of $(\beta, b, \tau, D^{-1} | y)$ as

$$\begin{aligned} p(\beta, b, \tau, D^{-1} | y) &\propto p(y | \beta, b, D) \pi(\beta | \tau) \pi(b | \tau, D) \pi(\tau) \pi(D^{-1}) \\ &\propto \tau^{p/2} \exp\left\{-\frac{\tau}{2}(y - X\beta - Zb)'(y - X\beta - Zb)\right\} \\ &\quad \times \exp\left\{-\frac{\tau}{2}(\beta - \mu_0)'\Sigma_0^{-1}(\beta - \mu_0)\right\} \\ &\quad \times |D^{-1}|^{\frac{n}{2}} \tau^{\frac{nq}{2}} \exp\left\{-\frac{\tau}{2}(b'(I_n \otimes D^{-1})b)\right\} \left(\tau^{\frac{\delta_0}{2}-1} \exp\left\{-\frac{\gamma_0\tau}{2}\right\}\right) \\ &\quad \times |D^{-1}|^{\frac{1}{2}(\nu_0-q-1)} \exp\left\{-\frac{1}{2}\text{tr}(C_0^{-1}D^{-1})\right\}. \end{aligned}$$

This joint posterior **does not** have an analytic closed form, but **all** of the complete conditionals have analytic closed forms. We now derive these complete conditional distributions.

(1) Let us first derive $p(\beta | y, b, \tau, D)$.

$$\begin{aligned} p(\beta | y, b, \tau, D) &\propto \exp \left\{ -\frac{\tau}{2}(y - X\beta - Zb)'(y - X\beta - Zb) \right\} \\ &\quad \times \exp \left\{ -\frac{\tau}{2}(\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \right\} \\ &\propto \exp \left\{ -\frac{\tau}{2}(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \right\} \exp \left\{ -\frac{\tau}{2}(\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \right\}, \end{aligned}$$

where

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

Thus

$$p(\beta | y, b, \tau, D) \propto \exp \left\{ -\frac{\tau}{2}(\beta - \beta^*)' (X'X + \Sigma_0^{-1})(\beta - \beta^*) \right\},$$

where

$$\beta^* = (X'X + \Sigma_0^{-1})^{-1} (X'X \hat{\beta} + \Sigma_0^{-1} \mu_0).$$

Thus

$$(\beta | y, b, \tau, D) \sim N_p(\beta^*, \tau^{-1}(X'X + \Sigma_0^{-1})^{-1}).$$

Remark 5.3

Since b is assumed known in $[\beta \mid y, b, \tau, D]$,

$$\begin{aligned}(y - X\beta - Zb)'(y - X\beta - Zb) &= (y - Zb - X\beta)'(y - Zb - X\beta) \\ &= (y^* - X\beta)'(y^* - X\beta),\end{aligned}$$

where $y^* = y - Zb$, and thus,

$$(y^* - X\beta)'(y^* - X\beta) = (y^* - \hat{y}^*)'(y^* - \hat{y}^*) + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}),$$

where

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y^* \\ &= (X'X)^{-1}X'(y - Zb),\end{aligned}$$

and

$$\hat{y}^* = X\hat{\beta}^* = X(X'X)^{-1}X'(y - Zb).$$

This is what facilitated an easy completion of the square in deriving $[\beta \mid y, b, \tau, D]$.

(2) Let us derive $[b \mid y, \beta, \tau, D]$.

$$\begin{aligned} p(b \mid y, \beta, \tau, D) &\propto \exp\left\{-\frac{\tau}{2}[y - X\beta - Zb]'[y - X\beta - Zb]\right\} \exp\left\{-\frac{\tau}{2}(b'(I_n \otimes D^{-1})b)\right\} \\ &\propto \exp\left\{-\frac{\tau}{2}(b - \hat{b})'Z'Z(b - \hat{b})\right\} \exp\left\{-\frac{\tau}{2}(b'(I_n \otimes D^{-1})b)\right\}, \end{aligned}$$

where

$$\hat{b} = (Z'Z)^{-1}Z'(y - X\beta).$$

Thus

$$p(b \mid y, \beta, \tau, D) \propto \exp\left\{-\frac{\tau}{2}(b - b^*)'(I_n \otimes D^{-1} + Z'Z)(b - b^*)\right\},$$

where

$$b^* = (I_n \otimes D^{-1} + Z'Z)^{-1}(Z'Z\hat{b}).$$

Thus

$$(b \mid y, \beta, \tau, D) \sim N_{nq}(b^*, \tau^{-1}(I_n \otimes D^{-1} + Z'Z)^{-1}).$$

Note that

$$I_n \otimes A_{q \times q} = \begin{pmatrix} A & & 0 \\ & \ddots & \\ 0 & & A \end{pmatrix}_{nq \times nq}.$$

(3) Let us derive $[\tau | y, \beta, b, D]$.

$$\begin{aligned} p(\tau | y, \beta, b, D) &\propto \tau^{\frac{N+p+nq+\delta_0}{2}-1} \\ &\times \exp \left\{ -\frac{\tau}{2} (y - X\beta - Zb)'(y - X\beta - Zb) + (\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) + \gamma_0 \right\} \\ &\times \exp \left\{ -\frac{\tau}{2} (b' (I_n \otimes D^{-1}) b) \right\}. \end{aligned}$$

Let

$$\begin{aligned} \gamma^* &= (y - X\beta - Zb)'(y - X\beta - Zb) + (\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \\ &\quad + \gamma_0 + b' (I_n \otimes D^{-1}) b, \end{aligned}$$

and $\delta^* = N + p + nq + \delta_0$. Then

$$[\tau | y, \beta, b, D] \sim \text{gamma} \left(\frac{\delta^*}{2}, \frac{\gamma^*}{2} \right).$$

(4) Finally, we must derive $[D^{-1} | y, \beta, b, \tau]$.

$$\begin{aligned}
 p(D^{-1} | y, \beta, b, \tau) &\propto \exp \left\{ -\frac{\tau}{2} (b' (I_n \otimes D^{-1}) b) \right\} | I_n \otimes D^{-1} |^{\frac{1}{2}} | D^{-1} |^{\frac{1}{2}(\nu_0 - q - 1)} \\
 &\quad \times \exp \left\{ -\frac{1}{2} \text{tr} (C_0^{-1} D^{-1}) \right\} \\
 &= | D^{-1} |^{\frac{n}{2}} | D^{-1} |^{\frac{1}{2}(\nu_0 - q - 1)} \exp \left\{ -\frac{1}{2} \text{tr} (C_0^{-1} D^{-1}) \right\} \\
 &\quad \times \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n (b_i' D^{-1} b_i) \right\} \\
 &= | D^{-1} |^{\frac{\nu_0 + n - q - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} (C_0^{-1} D^{-1}) \right\} \\
 &\quad \times \exp \left\{ -\frac{\tau}{2} \text{tr} \left(\sum_{i=1}^n (b_i' D^{-1} b_i) \right) \right\}.
 \end{aligned}$$

Now,

$$\begin{aligned}
 \text{tr} \left(\sum_{i=1}^n b_i' D^{-1} b_i \right) &= \sum_{i=1}^n \text{tr}(b_i' D^{-1} b_i) \\
 &= \sum_{i=1}^n \text{tr}(b_i b_i' D^{-1}) \\
 &= \text{tr} \left[\left(\sum_{i=1}^n b_i b_i' \right) D^{-1} \right].
 \end{aligned}$$

Thus

$$\begin{aligned}
 p(D^{-1} | y, \beta, b, \tau) &\propto |D^{-1}|^{\frac{\nu_0 + n - q - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left\{ (C_0^{-1} + \tau \Sigma b_i b_i') D^{-1} \right\} \right\} \\
 &= |D^{-1}|^{\frac{\nu_0 + n - q - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left\{ (C_0^{-1} + \tau B' B) D^{-1} \right\} \right\},
 \end{aligned}$$

where $B' = (b_1, \dots, b_n)_{q \times n}$, and therefore

$$B' B = \sum_{i=1}^n b_i b_i'.$$

We recognize this conditional posterior kernel as a Wishart. In particular,

$$[D^{-1} | y, \beta, b, \tau] \sim W_q(n + \nu_0, (C_0^{-1} + \tau B' B)^{-1}).$$

The full conditionals under the **noninformative priors** in (5.9) can be obtained by setting

$$\nu_0 = 0, \ C_0^{-1} = 0, \ \delta_0 = 0, \ \gamma_0 = 0, \ \Sigma_0^{-1} = 0.$$

In this case we obtain

(1) $[\beta \mid y, b, \tau, D] \sim N_p[\hat{\beta}, \tau^{-1}(X'X)^{-1}]$, where

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

(2) $[b \mid y, \beta, \tau, D] \sim N_{nq}(b^*, \tau^{-1}(I_n \otimes D^{-1} + Z'Z)^{-1})$, where

$$\begin{aligned} b^* &= (I_n \otimes D^{-1} + Z'Z)^{-1}(Z'Z\hat{\beta}) \\ \hat{b} &= (Z'Z)^{-1}Z'(y - X\beta). \end{aligned}$$

This is the same as the informative case.

(3) $[\tau \mid y, \beta, b, D] \sim \text{gamma}\left(\frac{\tilde{\delta}}{2}, \frac{\tilde{\gamma}}{2}\right)$, where

$$\begin{aligned} \tilde{\delta} &= p + nq \\ \tilde{\gamma} &= (y - X\beta - Zb)'(y - X\beta - Zb) + b'(I_n \otimes D^{-1})b. \end{aligned}$$

(4) $[D^{-1} \mid y, \beta, b, \tau] \sim W_q(n, \tau^{-1}(B'B)^{-1})$.

To implement the Gibbs sampler, we use the following steps:

- (0) start with initial values $(\beta^{(0)}, b^{(0)}, \tau^{(0)}, D^{-1(0)})$
- (1) generate $\beta^{(1)}$ from $[\beta | y, b^{(0)}, \tau^{(0)}, D^{-1(0)}]$
- (2) generate $b^{(1)}$ from $[b | y, \beta^{(1)}, \tau^{(0)}, D^{-1(0)}]$
- (3) generate $\tau^{(1)}$ from $[\tau | y, \beta^{(1)}, b^{(1)}, D^{-1(0)}]$
- (4) generate $D^{-1(1)}$ from $[D^{-1} | y, \beta^{(1)}, b^{(1)}, \tau^{(1)}]$
- (5) repeat steps (1) - (4) until we obtain M samples

$$\{(\beta^{(j)}, b^{(j)}, \tau^{(j)}, D^{-1(j)}), j = 1, \dots, M\}.$$

These samples will be samples from the joint posterior distribution of $(\beta, b, \tau, D^{-1} | y)$.

Remark 5.4

To generate a $W \sim W_q(n, \Sigma)$ matrix, we note that we can write

$$W = \sum_{i=1}^n y_i y_i' ,$$

where

$$y_i \stackrel{i.i.d.}{\sim} N_q(0, \Sigma), \quad i = 1, \dots, n.$$

Thus, all we need to do is generate n i.i.d. $N_q(0, \Sigma)$ random vectors

y_1, \dots, y_n , and compute $W = \sum_{i=1}^n y_i y_i'$. This gives us one $W_q(n, \Sigma)$ sample.

Remark 5.5

In general, introducing random effects into a model creates a correlation structure between the components of Y_i , marginally.

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i, \quad X_i \text{ is } n_i \times p, \quad Z_i \text{ is } n_i \times q,$$

$$Y_i|b_i \sim N(X_i\beta + Z_ib_i, \sigma^2 I_{n_i}), \quad i = 1, \dots, n.$$

Given b_i , the components of Y_i are independent.

Unconditionally, if $b_i \sim N_q(0, \sigma^2 D)$, then

$$Y_i \sim N(X_i\beta, \sigma^2(I_{n_i} + Z_i D Z_i')),$$

so now the components of Y_i are dependent.

$$\begin{aligned} p(y_i) &= \int p(y_i|b_i)\pi(b_i) db_i \\ &= N(X_i\beta, \sigma^2(I_{n_i} + Z_i D Z_i')) \end{aligned}$$

Generalized Linear Mixed Models

The generalized linear mixed model (GLMM) is the GLM generalization of the normal linear random effects model. It is defined as follows. Let y_{ij} denote the j^{th} measurement on the i^{th} subject. Suppose the sampling distribution of y_{ij} , $j = 1, \dots, n_i$, $i = 1, \dots, n$, is from an exponential family, so that

$$p(y_{ij} | \theta_{ij}, \phi) = \exp \left\{ \phi^{-1}(y_{ij}\theta_{ij} - a(\theta_{ij})) + c(y_{ij}, \phi) \right\} .$$

Without loss of generality, let us assume that $\phi = 1$ as in the logistic and Poisson regression models.

Thus we assume the y_{ij} 's are independent, each y_{ij} has canonical parameter θ_{ij} , and each y_{ij} comes from an exponential family. As before, we have

$$\begin{aligned}\mu_{ij} &= E(y_{ij} | \theta_{ij}) = a'(\theta_{ij}) = \frac{da(\theta_{ij})}{d\theta_{ij}} \\ v_{ij} &= \text{Var}(y_{ij} | \theta_{ij}) = a''(\theta_{ij}) = \frac{d^2a(\theta_{ij})}{d\theta_{ij}^2}.\end{aligned}$$

In the GLMM, the canonical parameter θ_{ij} is related to the covariates by

$$\theta(\theta_{ij}) = \eta_{ij} = x'_{ij}\beta + z'_{ij}b_i ,$$

where $\theta(\theta_{ij})$ is a monotonic function of θ_{ij} , x'_{ij} is a $1 \times p$ vector denoting the j^{th} row of X_i , and z'_{ij} is a $1 \times q$ vector denoting the j^{th} row of Z_i . Thus

$$X_i = \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{in_i} \end{pmatrix}_{n_i \times p}, \quad Z_i = \begin{pmatrix} z'_{i1} \\ \vdots \\ z'_{in_i} \end{pmatrix}_{n_i \times q}.$$

Again, the link function $\theta(\theta_{ij})$ is referred to as the θ -link, and $\eta_{ij} = x'_{ij}\beta + z'_{ij}b_i$ is the **linear predictor**. Assuming $\phi = 1$, we can write the GLMM as

$$p(y_{ij} | \beta, b_i) = \exp \left\{ y_{ij} \theta(x'_{ij}\beta + z'_{ij}b_i) - a(\theta(x'_{ij}\beta + z'_{ij}b_i)) + c(y_{ij}) \right\}.$$

Thus, **conditional** on the **random effects** b_i , the observations on subject i are independent. The likelihood function for all n subjects is thus given by

$$p(y | \beta, b) = \prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} | \beta, b_i),$$

where $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$. Each b_i is $q \times 1$ and β is $p \times 1$.

When $\theta(\theta_{ij}) = \theta_{ij} = \eta_{ij}$, then the link is a **canonical link**. For example, in the GLMM logistic regression model, we have

$$p(y_{ij} | \beta, b_i) = \exp \left\{ y_{ij}(x'_{ij}\beta + z'_{ij}b_i) - \log(1 + \exp(x'_{ij}\beta + z'_{ij}b_i)) \right\}.$$

Frequentist Likelihood-Based Inference for GLMM's

Likelihood based inference is based on the **marginal likelihood** with the random effects being integrated out. As usual, we assume

$$b_i \stackrel{i.i.d.}{\sim} N_q(0, D),$$

so that the full likelihood for subject i is

$$p(y_{ij} | \beta, b_i) \pi(b_i),$$

which leads to

$$p(y, b | \beta) = \prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} | \beta, b_i) \pi(b_i), \quad (5.15)$$

where

$$p(y_{ij} | \beta, b_i) = \exp \{ y_{ij} \theta(x'_{ij}\beta + z'_{ij}b_i) - a(\theta(x'_{ij}\beta + z'_{ij}b_i)) + c(y_{ij}) \}$$

and

$$\pi(b_i) = (2\pi)^{-q/2} |D|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} b'_i D^{-1} b_i \right\}.$$

The **marginal likelihood** of β is given by

$$p(y \mid \beta) = \int_{R^{nq}} p(y, b \mid \beta) \, db , \quad (5.16)$$

where $p(y, b \mid \beta)$ is given by (5.15).

We see that

$$\begin{aligned} p(y \mid \beta) &= \int_{R^{nq}} p(y, b \mid \beta) \, db \\ &= \int_{R^q} \cdots \int_{R^q} \left\{ \prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} \mid \beta, b_i) \pi(b_i) \, db_i \right\} \\ &= \prod_{i=1}^n \left[\int_{R^q} \prod_{j=1}^{n_i} p(y_{ij} \mid \beta, b_i) \pi(b_i) \, db_i \right] . \end{aligned} \quad (5.17)$$

Thus, by integrating out b_i , we induce a correlation structure **within subjects**, that is, $(y_{i1}, \dots, y_{in_i})$ become correlated.

Thus the marginal likelihood involves evaluating n q -dimensional integrals. That is, (5.17) involves n q -dimensional integrals over R^q . For the general class of GLMM's, these integrals do not have a closed form and are very difficult to evaluate. Thus frequentist likelihood-based inference from the GLMM is **essentially impossible**, since the marginal likelihood cannot be computed or even well approximated.

This problem led to the development of non-likelihood based methods, and, in particular, methods based on **Generalized Estimating Equations** (GEE). These methods are quite common among frequentists and have been made popular by Zeger & Liang (1986, *Biometrics*) and Liang and Zeger (1986, *Biometrics*). See Diggle, Liang, and Zeger (1994, Oxford University Press) for a good discussion of GEE methods.

GEE's can be viewed as multivariate analogs of quasi-likelihood (Wedderburn, 1974, *Biometrika*). The main idea of GEE's is to develop a set of **score equations** or **estimating equations** to solve for (β, D) .

The estimating equations take the form

$$S_\beta = \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)' \text{Var}(y_i)^{-1} (y_i - \mu_i) = 0 ,$$

where

$$\text{Var}(y_i) \equiv \text{Var}(y_i | \beta)$$

and

$$\mu_i = E(y_i | \beta) .$$

Since $E(y_i | \beta)$ and $\text{Var}(y_i | \beta)$ depend on D , we replace D by a consistent estimator, \hat{D} , and solve $S_\beta(\beta, \hat{D}) = 0$. The covariance matrix D may be estimated by simultaneously solving $S_\beta(\beta, D) = 0$ and

$$S_D(\beta, D) = \sum_{i=1}^n \left(\frac{\partial \tilde{\mu}_i}{\partial D} \right)' H_i^{-1} (w_i - \tilde{\mu}_i) = 0 ,$$

where

$$w_i = (y_{i1}y_{i2}, y_{i1}y_{i3}, \dots, \dots, y_{in_{i-1}}y_{in_i}, y_{i1}^2, \dots, y_{n_i}^2)'$$

is the set of all products of pairs of response and squared responses, and $\tilde{\mu}_i = E(w_i | \beta)$.

The choice of the weight matrix, H_i , depends on the type of responses. For binary responses, the last n_i components of w_i can be ignored since the variance of a binary response is determined by its mean. In this case,

$$H_i = \begin{bmatrix} \text{Var}(y_{i1}y_{i2} | \beta) & & 0 \\ & \text{Var}(y_{i1}y_{i3} | \beta) & \\ & & \ddots \\ 0 & & \text{Var}(y_{in_{i-1}}y_{in_i} | \beta) \end{bmatrix}.$$

See Diggle, Liang, and Zeger for a detailed account of GEE methods. GEE methods have been implemented in *Proc MIXED* and *Proc GLIMMIX* in SAS.

Bayesian Methods for GLMM's

Likelihood-based Bayesian methods are relatively straightforward to carry out using the Gibbs sampler. Due to the nature of the Gibbs sampler, we do **not** need to integrate the b_i 's for making inferences about β . Thus, no intractable integrals arise for Bayesian inference for GLMM's. Thus, for this class of models, Bayesian methods are **much more** powerful than frequentist methods, in which likelihood-based inference is **not** even possible due to the integrals.

In the Bayesian approach, we write the kernel of the **joint posterior** of (β, b, D) and then run the Gibbs sampler on the complete conditionals. The complete conditionals **do not** have an analytic closed form for the original parameterization of the GLMM, so rejection algorithms must be used within each Gibbs cycle.

Prior Distribution for (β, D^{-1})

The usual **joint** proper prior for (b, β, D^{-1}) is

$$\pi(\beta, b, D^{-1}) = \pi(\beta) \pi(b | D^{-1}) \pi(D^{-1}) .$$

We take

$$\begin{aligned}\beta &\sim N_p(\mu_0, \Sigma_0) \\ D^{-1} &\sim W_q(\nu_0, C_0) \\ b | D &\sim N_{nq}(0, (I_n \otimes D)) .\end{aligned}$$

We note again that we have assumed $\phi = 1$ (i.e., $\tau = 1$) for GLMM's.

Thus the **joint posterior** of (β, b, D^{-1}) can be written as

$$p(\beta, b, D^{-1}) \propto \left\{ \prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} | \beta, b_i) \pi(b_i) \right\} \pi(\beta) \pi(D^{-1}),$$

where

$$\begin{aligned}\pi(b_i) &= (2\pi)^{-q/2} |D|^{-1/2} \exp \left\{ -\frac{1}{2} b_i' D^{-1} b_i \right\} \\ \pi(\beta) &= (2\pi)^{-p/2} |\Sigma_0|^{-1/2} \exp \left\{ -\frac{1}{2} (\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \right\} \\ \pi(D^{-1}) &\propto |D^{-1}|^{\frac{1}{2}(\nu_0 - q - 1)} \exp \left\{ -\frac{1}{2} \text{tr}(C_0^{-1} D^{-1}) \right\}.\end{aligned}$$

To do Gibbs sampling, we need to write out the complete conditionals up to a normalizing constant, and then use the adaptive rejection algorithm to sample from each complete conditional.

For the GLMM, the complete conditionals can be written as follows:

1)

$$p(\beta_j \mid y, b, \beta_i, D, i \neq j) \propto \left(\prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} \mid \beta, b_i) \right) \pi(\beta), \quad j = 1, \dots, p,$$

where

$$p(y_{ij} \mid \beta, b_i) = \exp \{ y_{ij}(x'_{ij}\beta + z'_{ij}b_i) - \log(1 + \exp(x'_{ij}\beta + z'_{ij}b_i)) \} .$$

Conditional on (b_i, y) , each $[\beta_j \mid y, b, \beta_i, i \neq j]$ is log-concave in β_j and hence ARS can be used to sample from $[\beta_j \mid y, b, \beta_i, i \neq j]$.

2)

$$p(b_{ij} \mid y, \beta, D, b_{i'j'}, (i', j') \neq (i, j)) \propto \prod_{j=1}^{n_i} p(y_{ij} \mid \beta, b_i) \pi(b_i) ,$$

where $\pi(b_i) \propto \exp \left\{ -\frac{1}{2} b_i' D^{-1} b_i \right\}$.

Again, we can use ARS to sample from $[b_{ij} \mid y, \beta, D, b_{i'j'}, (i', j') \neq (i, j)]$.

3)

$$\begin{aligned}
 p(D^{-1} | y, b, \beta) &\propto \pi(b) \pi(D^{-1}) \\
 &\propto |I_n \otimes D^{-1}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2} b'(I_n \otimes D^{-1}) b\right\} \\
 &\quad \times |D^{-1}|^{\frac{v_0-q-1}{2}} \exp\left\{-\frac{1}{2} \text{tr}(C_0^{-1} D^{-1})\right\} \\
 &= |D^{-1}|^{\frac{v_0+n-q-1}{2}} \exp\left\{-\frac{1}{2} \text{tr}\left((C_0^{-1} + B'B)D^{-1}\right)\right\}.
 \end{aligned}$$

Thus $[D^{-1} | y, \beta, b] = W_q(v_0 + n, (C_0^{-1} + B'B)^{-1})$.

Remarks 5.6

- We see that step 2) does not depend on all of the subjects, but only the i^{th} subject.
- For the posterior of $[\beta, b, D^{-1} | y]$ to be proper, the prior on D^{-1} must be proper.
- An improper prior on β , i.e., $\pi(\beta) \propto 1$, can still result in a proper posterior for (β, b, D^{-1}) .

For more on Gibbs sampling approaches to GLMM's, see Zeger and Karim (1991, *JASA*).

Efficient Parameterizations for GLMM's: Hierarchical Centering

Due to the fact that there are an inherently large number of parameters in the GLMM due to the random effects, **high correlations** can result in the parameters, thus creating potential convergence problems for the Gibbs Sampler. This problem is quite frequent in random effects models in general. One way to resolve this issue is to reparameterize the model.

Gelfand, Sahu, Carlin (1996, *Bayesian Statistics 5*) propose a very useful parameterization technique called **hierarchical centering**, and seems to be quite powerful in improving convergence for the class of GLMM's.

To demonstrate the idea of hierarchical centering, consider the one-way ANOVA model with random effects,

$$y_i = \mu + \alpha_i + \epsilon_i, \quad i = 1, \dots, m, \quad (5.18)$$

where $\epsilon_i \sim N(0, \sigma_e^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$, and $\mu \sim N(\mu_0, \sigma_\mu^2)$.

The model in (5.18) can be written in **hierarchical form**. Defining

$$\eta_i = \mu + \alpha_i ,$$

we have

$$y_i | \eta_i \sim N(\eta_i, \sigma_e^2),$$

$$\eta_i | \mu \sim N(\mu, \sigma_\alpha^2),$$

$$\mu \sim N(\mu_0, \sigma_\mu^2).$$

We refer to the transformation from $(\alpha_1, \dots, \alpha_m)$ to (η_1, \dots, η_m) as **hierarchical centering**. The main idea behind this reparameterization is that it results in a better behaved likelihood or posterior surface. From this one-way ANOVA model, it is straightforward to show that the joint posteriors of $(\mu, \alpha_1, \dots, \alpha_m)$ and $(\mu, \eta_1, \dots, \eta_m)$ are multivariate normal.

A comparison of posterior covariance matrices will tell us how “well-behaved” the posterior densities are. In particular, it can be shown that

$$\text{Corr}(\eta_i, \mu | y) = \left(1 + \frac{b\sigma_\alpha^2}{\sigma_e^2 \sigma_\mu^2}\right)^{-\frac{1}{2}},$$

$$\text{Corr}(\eta_i, \eta_j | y) = \left(1 + \frac{b\sigma_\alpha^2}{\sigma_e^2 \sigma_\mu^2}\right)^{-1},$$

where $b = \sigma_e^2 + \sigma_\alpha^2 + m\sigma_\mu^2$.

The limiting behavior of the correlations above as the variance components are made large is readily studied. For example, if $\sigma_\alpha^2 \rightarrow \infty$, then for fixed σ_e^2 and σ_μ^2 these correlations tend to zero. This is also true, if, in addition, we let $\sigma_\mu^2 \rightarrow \infty$. On the other hand, if $\sigma_e^2 \rightarrow \infty$, the correlations do not approach 0, and in fact will tend to 1 if $\sigma_\mu^2 \rightarrow \infty$ as well. In the original parameterization, we have

$$\text{Corr}(\alpha_i, \mu | y) = -\left(1 + \frac{b\sigma_e^2}{\sigma_\alpha^2 \sigma_\mu^2}\right)^{-\frac{1}{2}},$$

$$\text{Corr}(\alpha_i, \alpha_j | y) = \left(1 + \frac{b\sigma_e^2}{\sigma_\alpha^2 \sigma_\mu^2}\right)^{-1}.$$

These correlations tend to 0 as $\sigma_e^2 \rightarrow \infty$, but do **not** approach 0 as $\sigma_\alpha^2 \rightarrow \infty$, and in fact tend to 1 if $\sigma_\mu^2 \rightarrow \infty$ as well. In practice, when the random effects are needed, the error variance is much reduced. Thus σ_e^2 will rarely dominate the variability, so that the centered parameterization will likely be preferred.

Intuition may be attached as follows. Data y_i informs directly about $\eta_i = \mu + \alpha_i$. If, a priori, we know little about η_i , i.e., σ_α^2 is large, we will be unable to separate μ from α_i . Large μ implies small α_i , and vice-versa, so μ and α_i are, a posteriori, highly negatively correlated, and thus α_i and α_j are highly positively correlated.

We have assumed here that σ_e^2 is known for simplicity. With the variance components known, i.e., σ_e^2 is assumed known, the decision as to whether or not to center is **not** data driven, but rather emerges from the modeling specification. In applications, however, these components will be unknown, so we should instead consider the **joint posterior distribution** of $(\mu, \alpha, \sigma_\mu^2, \sigma_\alpha^2, \sigma_e^2 | y)$.

For GLMM's, we use the same idea, and the hierarchical centering parameterization involves taking

$$\eta_{ij} = x'_{ij}\beta + z'_{ij}b_i$$

$$j = 1, \dots, n_i, i = 1, \dots, n.$$

We then consider the joint posterior of $(\beta, \eta, D^{-1} | y)$, where $\eta = (\eta_{11}, \eta_{12}, \dots, \eta_{nn_n})'$, and η is $N \times 1$. We note that with this reparameterization the complete conditionals are

1. $[\eta | \beta, y, D]$ (log-concave, use ARS)
2. $[\beta | \eta, y, D]$ (normal if $\pi(\beta)$ is normal or uniform)
3. $[D^{-1} | \beta, \eta, y]$ (Wishart, same as before)

Note that $[\eta | \beta, y, D]$ does **not** have a closed form, but the complete conditionals of

$$[\eta_{ij} | \beta, y, D, \eta_{i'j'}, (i'j') \neq (i,j)]$$

are log-concave and hence ARS is easily applied.

We also note that $[\beta | \eta, y, D]$ is normal if $\pi(\beta)$ is normal or $\pi(\beta)$ is a uniform improper prior.

Thus, the hierarchical centering makes the complete conditionals a bit easier to sample from and also reduces the posterior correlations between the random effects.

Exercise 5.3

Derive $[\beta | \eta, y, D]$.

Example 5.3: Random Effects Logistic Regression with Seed Data

This example is taken from Crowder (1978), and concerns the proportion of seeds that germinated on each of 21 plates arranged according to a 2×2 factorial layout by seed and type of root extract. The data are shown below, where r_i and n_i are the number of germinated and the total number of seeds on the i^{th} plate, $i = 1, \dots, N$. These data are also analysed by, for example, Breslow and Clayton (1993).

seed <i>O. aegyptiaco</i> 75			seed <i>O. aegyptiaco</i> 73								
Bean			Cucumber			Bean			Cucumber		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	.26	5	6	.83	8	16	.50	3	12	.25
23	62	.37	53	74	.72	10	30	.33	22	41	.54
23	81	.28	55	72	.76	8	28	.29	15	30	.50
26	51	.51	32	51	.63	23	45	.51	32	51	.63
17	39	.44	46	79	.58	0	4	.00	3	7	.43
			10	13	.77						

The model is essentially a random effects logistic, allowing for over-dispersion. If p_i is the probability of germination on the i^{th} plate, we assume

$$\begin{aligned} r_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &= \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i \\ b_i &\sim N(0, \sigma^2), \end{aligned}$$

where x_{1i} and x_{2i} are the seed type and root extract of the i^{th} plate, respectively, and an interaction term $\alpha_{12}x_1x_2$ is included.

Let α be the vector of regression parameters and $\tau = \frac{1}{\sigma^2}$. We specify the priors to be

$$\begin{aligned} \alpha &\sim N_4(0, 100^2 I), \\ \tau &\sim \text{gamma}(.001, .001). \end{aligned}$$

The graphical model is shown in Figure 5.1.

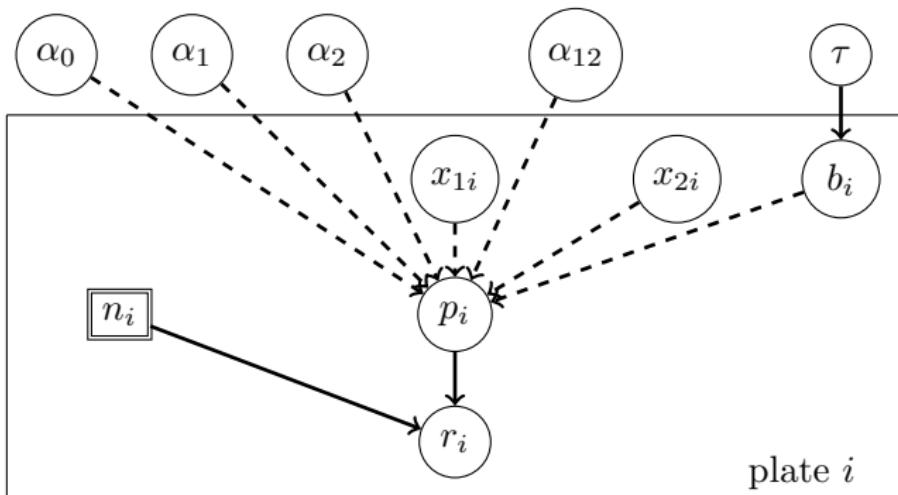


Figure 5.1: Graphical model for seeds example

The log-likelihood for an observation r_i arising from a binomial model with denominator n_i and success probability p_i is

$$\ell(p_i) = r_i \log(p_i) + (n_i - r_i) \log(1 - p_i).$$

The saturated log-likelihood for the binomial model is

$$\ell_{sat}(p_i) = r_i \log\left(\frac{r_i}{n_i}\right) + (n_i - r_i) \log\left(1 - \frac{r_i}{n_i}\right).$$

Hence the deviance is given by

$$2 \left(\sum_i \ell_{sat}(p_i) - \sum_i \ell(p_i) \right).$$

To fit the model in JAGS, we first reformat the data from binomial form into stretched form by creating an observation for each seed.

For the j^{th} seed on the i^{th} plate, we let y_{ij} be the indicator if the seed germinated, where $y_{ij} \sim \text{Bernoulli}(p_i)$, $i = 1, \dots, N$, $j = 1, \dots, n_i$.

Data in stretched form can be written as

Obs	y	x1_st	x2_st	plate
1	1	0	0	1
2	1	0	0	1
3	1	0	0	1
.....				
.....				
830	0	1	1	21
831	0	1	1	21

The deviance for this model can be calculated within JAGS using the data in the original binomial form.

We write out the model and calculate the deviance in the file *seedJAGS.txt*, as shown below:

```

model{
  for(i in 1:Nobs){
    y[i] ~ dbern(prob[i])
    logit(prob[i]) = alpha[1] + x1_st[i]*alpha[2] + x2_st[i]*alpha[3] + x1_st[i]*x2_st[i]*alpha[4] + b[plate[i]]
  }
  # Specify precision (not variance or standard deviation) in normal distribution
  for(j in 1:N){
    b[j] ~ dnorm(0, tau)
  }

  # Prior distributions
  for(i in 1:4){
    alpha[i] ~ dnorm(0, 0.0001)
  }
  tau ~ dgamma(0.001, 0.001)

  # Calculate deviance
  for(i in 1:N){
    logit(p[i]) = alpha[1] + x1[i]*alpha[2] + x2[i]*alpha[3] + x1[i]*x2[i]*alpha[4] + b[i]
    l_p[i] = r[i]*log(p[i]) + (n[i] - r[i])*log(1 - p[i])
    l_sat_p[i] = r[i]*log(r[i]/n[i]) + (n[i] - r[i])*log(1 - r[i]/n[i])
  }
  dev = 2 * (sum(l_sat_p) - sum(l_p))
}

```

R code:

```

# Read in original data in binomial form
r <- c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0, 3, 22, 15, 32, 3)
n <- c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45, 4, 12, 41, 30, 51, 7)
x1 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
x2 <- c(0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1)
N <- length(n)           # number of plates

# Reformat data into stretched form
Nobs <- sum(n)          # total number of observations
y <- NULL                # indicator if individual seed germinated
for(i in 1:length(n)){
  y <- append( y, rep( c(1,0), c(r[i], n[i] - r[i]) ), after = length(y) )
}
x1_st <- rep(x1, n)      # seed type, stretched
x2_st <- rep(x2, n)      # root extract type, stretched
plate <- rep(1:N, n)      # plate labels

# Prepare input data for JAGS model
data.jags <- list( y = y, x1_st = x1_st, x2_st = x2_st, Nobs = Nobs, N = N,
                   r = r, n = n, x1 = x1, x2 = x2, plate = plate )

# Set up model using JAGS ("rjags" package)
library(rjags)
set.seed(779)
seed.inits <- list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 779)
jags.mod <- jags.model( file = "seedJAGS.txt", data = data.jags, inits = seed.inits,
                        n.adapt = 1000, n.chains = 1 )

```

```
# Specify names of parameters and deviance to sample
jags.vars <- numeric(5)
for(i in 1:4) jags.vars[i] <- paste( "alpha[", i, "]", sep = "" )
jags.vars[5] <- "dev"
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )
```

```
#Summary of parameter estimates and posterior deviance
> summary( jags.samps )
```

Iterations = 1001:11000

Thinning interval = 1

Number of chains = 1

Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
alpha[1]	-0.53759	0.1899	0.001899	0.009074
alpha[2]	0.05328	0.3204	0.003204	0.014207
alpha[3]	1.33908	0.2641	0.002641	0.011665
alpha[4]	-0.79444	0.4311	0.004311	0.019907
dev	25.42650	6.9065	0.069065	0.300131

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
alpha[1]	-0.8996	-0.6630	-0.54640	-0.4198	-0.13795
alpha[2]	-0.5990	-0.1559	0.06398	0.2673	0.65870
alpha[3]	0.8084	1.1658	1.34345	1.5105	1.85652
alpha[4]	-1.6670	-1.0732	-0.79172	-0.5099	0.04993
dev	13.4920	20.2996	24.87752	30.3183	39.12985

Analysis

We may compare simple logistic, maximum likelihood (from EGRET), penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993), and JAGS results using a burn-in of 1000 iterations and estimation based on 10,000 samples.

variable	Logistic regression	maximum likelihood	PQL	JAGS
	$\beta \pm SE$	$\beta \pm SE$	$\beta \pm SE$	$\beta \pm SE$
constant (α_0)	-.558 ± .126	-.548 ± .167	-.542 ± .190	-.538 ± .190
seed (α_1)	.146 ± .223	.097 ± .278	.077 ± .308	.053 ± .320
extract (α_2)	1.318 ± .177	1.337 ± .237	1.339 ± .270	1.339 ± .264
interaction (α_{12})	-.778 ± .306	-.811 ± .385	-.825 ± .430	-.794 ± .431
scale (σ)	—	.236 ± .110	.313 ± .121	.287 ± .142

JAGS produces samples for the deviance just like any other parameter. Hence we obtain a *distribution* for the deviance as shown in Figure 5.2.

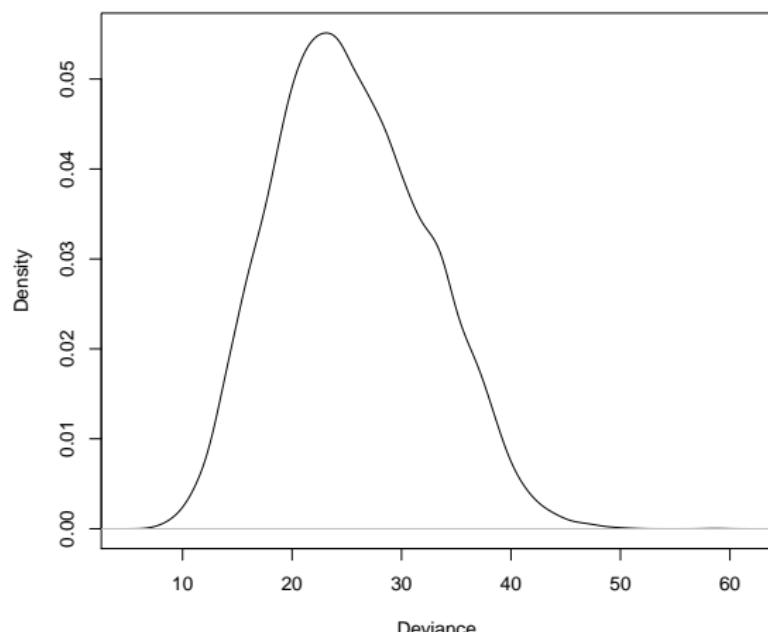


Figure 5.2: Posterior distribution of the deviance for the seeds example

Example 5.4: Institutional Ranking with Surgical Data

This example considers mortality rates in 12 hospitals performing cardiac surgery in babies. The data are shown below.

	A	B	C	D	E	F	G	H	I	J	K	L
No. of ops. (n)	47	148	119	810	211	196	148	215	207	97	256	360
No. of deaths (r)	0	18	8	46	8	13	9	31	14	8	29	24

Let n_i be the number of operations for hospital i . The number of deaths r_i are modelled as a binary response variable with “true” failure probability p_i , i.e.,

$$r_i \sim \text{Binomial}(n_i, p_i), \quad i = 1, \dots, 12.$$

We will fit both a fixed effects model and a random effects model and compare the two models.

(i) Fixed effects model

We first assume that the true failure probabilities are *independent* (i.e. fixed effects) for each hospital. This is equivalent to assuming a standard non-informative prior distribution for the p_i 's, namely

$$p_i \sim \text{Beta}(1.0, 1.0).$$

The Stan code from the file *surgery-fe.stan* for fixed effects model is below:

```
data {  
    // Define variables in data  
    int<lower=0> N;                      // Number of hospitals  
    int<lower=0> n[N];                    // Number of operations per hospital  
    int<lower=0> r[N];                    // Number of deaths per hospital  
}  
  
parameters {  
    // Define parameters to estimate  
    real<lower=0, upper=1> p[N];  
}  
  
model {  
    // Prior part of Bayesian inference  
    p ~ beta(1,1);  
  
    // Likelihood part of Bayesian inference  
    r ~ binomial(n, p);  
}
```

R code:

```
# Read in data
r <- c(0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24)
n <- c(47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360)
N <- 12
surgery <- list( N = N, r = r, n = n )

# Run Stan code
library(rstan)
fileName.fe <- "./surgery_fe.stan"
stan.code.fe <- readChar(fileName.fe, file.info(fileName.fe)$size)
stan.mod.fe <- stan( model_code = stan.code.fe, data = surgery, chains = 4, iter = 3500,
                      warmup = 1000, thin = 1, control = list(max_treedepth = 15) )

# Calculate posterior means and 95% credible intervals
params.fe <- as.matrix(stan.mod.fe)
hosp.means.fe <- colMeans(params.fe[,1:N])
low.lims.fe <- apply( params.fe[,1:N], 2, quantile, .025 )
up.lims.fe <- apply( params.fe[,1:N], 2, quantile, .975 )
```

(ii) Random effects model

A more realistic model for the surgical data is to assume that the failure rates across hospitals are *similar* in some way. This is equivalent to specifying a *random effects* model for the true failure probabilities p_i , as follows:

$$\begin{aligned}\text{logit}(p_i) &= b_i, \\ b_i &\sim N(\mu, \sigma^2).\end{aligned}$$

Let $\tau = \frac{1}{\sigma^2}$. We specify the following standard non-informative priors for the population mean (logit) probability of failure, μ , and precision, τ :

$$\begin{aligned}\mu &\sim N(0, 100^2), \\ \tau &\sim \text{gamma}(.001, .001).\end{aligned}$$

Figure 5.3 shows the graph corresponding to the above model.

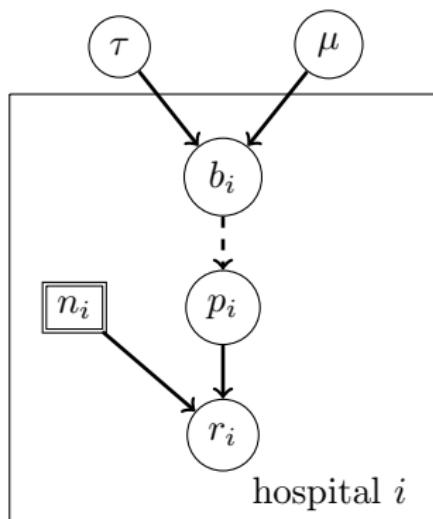


Figure 5.3: Graphical model for the random effects surgical example

We write out the random effects model in the file *surgery-re.stan*, as shown below:

```

data {
    // Define variables in data
    int<lower=0> N;                      // Number of hospitals
    int<lower=0> n[N];                   // Number of operations per hospital
    int<lower=0> r[N];                   // Number of deaths per hospital
}

parameters {
    // Define parameters to estimate
    vector[N] b;
    real<lower=0> tau;
    real mu;
}

transformed parameters {
    real<lower=0> sigma2b;
    sigma2b = 1/tau;
}

model {
    // Random effects
    // Enter stand. dev. (not variance) in normal distribution matrix
    b ~ normal(mu, sqrt(sigma2b));

    // Prior part of Bayesian inference
    mu ~ normal(0, 100);
    tau ~ gamma(.001, .001);

    // Likelihood part of Bayesian inference
    for (i in 1:N) {
        r[i] ~ binomial(n[i], inv_logit(b[i]));
    }
}

```

R code:

```
# Read in data
r <- c(0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24)
n <- c(47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360)
N <- 12
surgery <- list( N = N, r = r, n = n )

# Run Stan code
library(rstan)
fileName.re <- "./surgery_re.stan"
stan.code.re <- readChar(fileName.re, file.info(fileName.re)$size)
stan.mod.re <- stan( model_code = stan.code.re, data = surgery,
                     chains = 4, iter = 14000, warmup = 1000, thin = 5,
                     control = list(max_treedepth = 15, adapt_delta = 0.999) )
params.re <- as.matrix(stan.mod.re)
probs.re <- matrix( 0, nrow = nrow(params.re), ncol = N )
for(i in 1:N){
  probs.re[,i] <- exp( params.re[,i] ) / (1 + exp( params.re[,i] ))
}

# Calculate posterior means and 95% credible intervals
hosp.means.re <- colMeans(probs.re)
low.lims.re <- apply( probs.re, 2, quantile, .025 )
up.lims.re <- apply( probs.re, 2, quantile, .975 )
```

Figure 5.4 shows the posterior mean and 95% credible interval for the estimated surgical mortality rate in each hospital for both the fixed and random effect models.

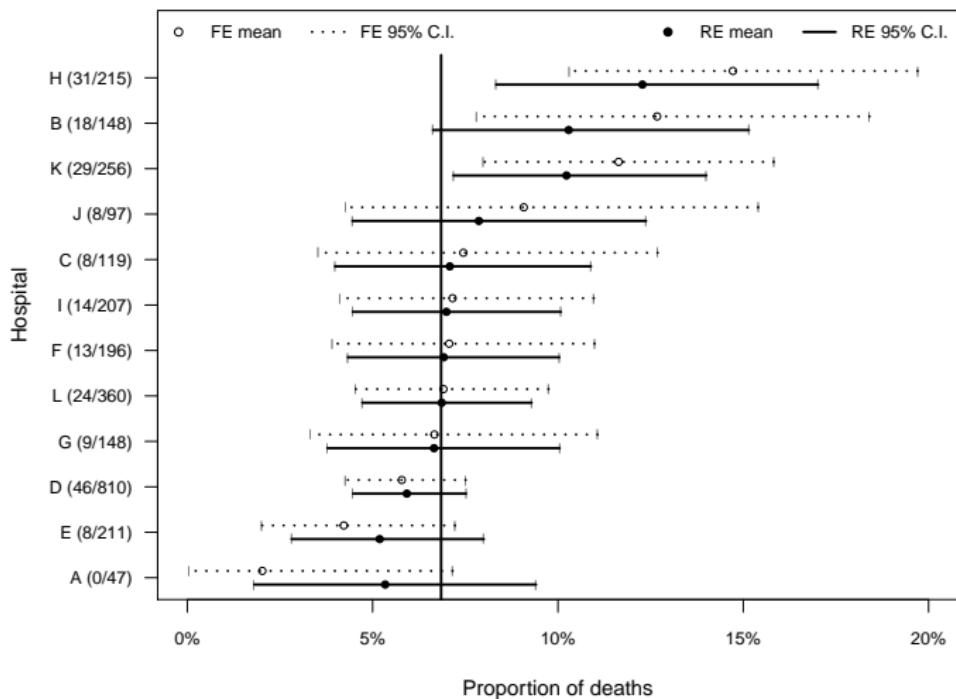


Figure 5.4: Fixed and shrunk estimates of the surgical mortality rates in each hospital. Numbers in brackets show the observed number of deaths and the total number of operations. The vertical line at $p = 6.8\%$ indicates the population mean failure rate estimated from the random effects model.

(iii) Ranking each hospital

A particular strength of the Markov chain Monte Carlo (Gibbs sampling) approach implemented in Stan and JAGS is the ability to make inferences on arbitrary functions of unknown model parameters. For example, we may compute the *rank* probability of failure for each hospital at each iteration. This yields a sample from the posterior distribution of the ranks which may be summarized to provide an estimate of the mean or median rank for each hospital, plus a 95% credible interval. The latter captures the (typically large) uncertainty associated with the rank position of each hospital.

We compute the ranks in Stan using the `step` function. To do so, we add the following “generated quantities” program block to the end of the code in the file `surgery_re.stan`, and then we run the same R code used for the random effects model:

```
generated quantities {
  real<lower=0, upper=1> p[N];
  matrix[N,N] not_less_than;
  real<lower=0> hosp_rank[N];

  for (i in 1:N) {
    p[i] = inv_logit(b[i]);
  }
  for (i in 1:N) {
    for (j in 1:N) {
      not_less_than[i,j] = step (p[i] - p[j]);
    }
    hosp_rank[i] = sum(not_less_than[i,]);
  }
}
```

The function `step` (x) = 1 if $x \geq 0$ and 0 otherwise. The i^{th} row of the array `not_less_than` thus contains a 1 in columns corresponding to hospitals with an equal or lower estimated failure probability than hospital i , and zeros elsewhere. Summing this row yields the total number of hospitals who have a “better” (lower) failure rate than hospital i , and thus corresponds to that hospital’s rank.

Results

Figure 5.5 shows the posterior mean and 95% credible interval for the estimated surgical mortality rate in each hospital for both the fixed and random effect models. These interval estimates illustrate the considerable uncertainty associated with “league tables”: there are only 2 hospitals (H and K) whose intervals exclude the median rank and none whose intervals fall completely within the lower or upper quartiles.

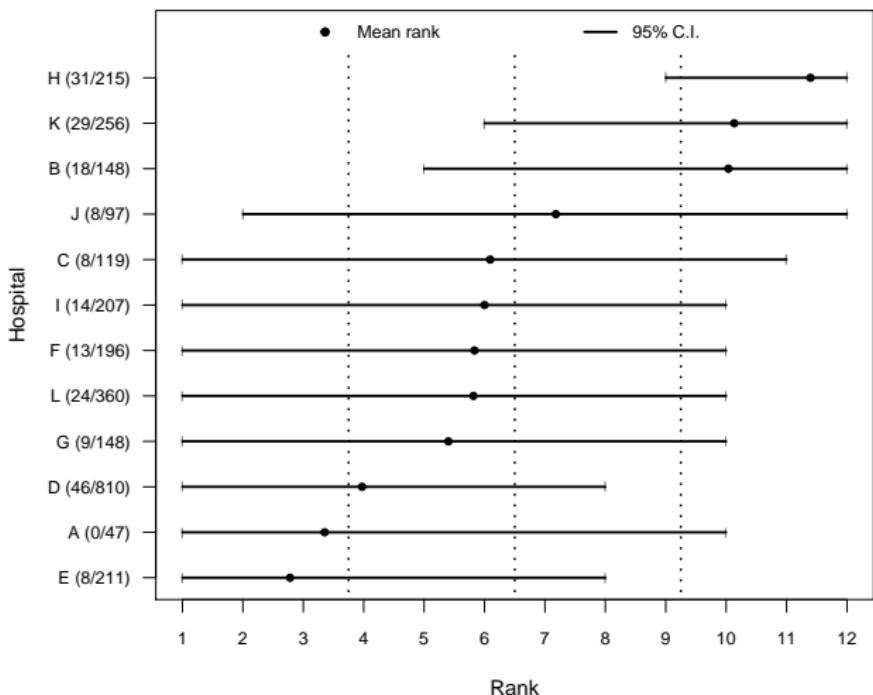


Figure 5.5: Posterior means and 95% credible intervals for the rank of each hospital. Vertical dashed lines indicate the position of lower and upper quartiles and median rank.

Example 5.5: Repeated Measures on Poisson counts

We now consider an example with data from a clinical trial of 59 epileptics. (Thall & Vail, 1990 Biometrics).

Patients suffering from simple or complex partial seizures were randomized to receive either an epileptic drug or placebo, or an adjuvant to chemotherapy. At each of 4 clinic visits (post randomization), the number of seizures occurring over the previous 2 weeks was reported. We denote this as Y_{jk} for the j^{th} patient and the k^{th} visit.

The covariates we look at include treatment (1 for treatment, 0 for placebo), the log of the baseline eight week seizure count (before randomization), the interaction between treatment and the log of the baseline count, the log of age, and V4 (0/1 indicator variable of visit 4). We center the log of the baseline count and the log of age.

We fit the model using both Stan and BGLIMM in SAS.

We consider the model $y_{jk} \sim \text{Poisson}(\mu_{jk})$, where

$$\begin{aligned}\log(\mu_{jk}) &= \alpha_0 + \alpha_{base} \log(Base_j/4) + \alpha_{trt} Trt_j + \alpha_{BT} Trt_j \log(Base_j/4) \\ &\quad + \alpha_{age} \log(Age_j) + \alpha_{V4} V4_k + b1_j + b_{jk},\end{aligned}$$

$$\begin{aligned}b1_j &\sim N(0, \sigma_{b1}^2), \\ b_{jk} &\sim N(0, \sigma_b^2),\end{aligned}$$

where $b1_j$ is a subject random effect and b_{jk} is a subject-by-visit random effect, $j = 1, \dots, 59$, $k = 1, \dots, 4$.

Let α be the vector of regression parameters, $\tau_{b1} = \frac{1}{\sigma_{b1}^2}$, and $\tau_b = \frac{1}{\sigma_b^2}$. We specify the priors to be

$$\begin{aligned}\alpha &\sim N_6(0, 100^2 I), \\ \tau_{b1} &\sim \text{gamma}(.001, .001), \\ \tau_b &\sim \text{gamma}(.001, .001).\end{aligned}$$

We write out the random effects model in the file *epilepsy.stan*, as shown below:

```
data {  
    // Define variables in data  
    int<lower=0> N;                      // Number of patients  
    int<lower=0> Nobs;                     // Number of observations  
    int<lower=0> K;                      // Number of visits  
    int<lower=0> y[Nobs];                  // Count outcomes  
    int<lower=0> p;                      // Number of fixed effects  
    matrix[Nobs,p] X;                     // Design matrix  
    int sub[Nobs];                      // Subject indicator  
}  
  
parameters {  
    // Define parameters to estimate  
    vector[p] alpha;  
    vector[N] b1;                         // Subject random effects  
    vector[Nobs] b;                       // Subject-by-visit random effects  
    real<lower=0> taub1;  
    real<lower=0> taub;  
}
```

```
transformed parameters {
  real<lower=0> sigma2b1;
  real<lower=0> sigma2b;
  real lp[Nobs];
  real <lower=0> mu[Nobs];

  sigma2b1 = 1/taub1;
  sigma2b = 1/taub;

  for (i in 1:Nobs) {
    // Linear predictor
    lp[i] = X[i,] * alpha + b1[sub[i]] + b[i];
    mu[i] = exp(lp[i]);
  }
}

model {
  // Random effects
  // Entering scalar as stand. dev. in multivariate normal distribution produces diagonal matrix
  b1 ~ normal(0, sqrt(sigma2b1));
  b ~ normal(0, sqrt(sigma2b));

  // Prior part of Bayesian inference
  alpha ~ normal(0, 100);
  taub1 ~ gamma(.001, .001);
  taub ~ gamma(.001, .001);

  // Likelihood part of Bayesian inference
  for (i in 1:Nobs) {
    y[i] ~ poisson(mu[i]);
  }
}
```

R code:

```
# Read in data and center specific covariates
epil <- read.table("epilepsy.dat", header = TRUE)
N <- length( unique(epil$subject) )
Nobs <- nrow(epil)
logBase4 <- log(epil$base / 4)
trt.lB4 <- epil$trt * logBase4
X <- data.frame( int = 1, logBase4 = logBase4 - mean(logBase4), trt = epil$trt,
                  trt.logBase4 = trt.lB4, logAge = log(epil$age) - mean(log(epil$age)),
                  V4 = epil$V4 )

# Arrange data for Stan
dat <- list()
dat$N <- N                      # number of patients
dat$Nobs <- Nobs                 # number of observations
dat$K <- 4                        # number of visits
dat$y <- epil$y                   # Count outcomes
dat$X <- X                        # design matrix (centered)
dat$p <- 6                        # number of fixed effects
dat$sub <- epil$subject           # subject indicators

# Run Stan code
library(rstan)
file.stan <- "epilepsy.stan"
stan.code <- readChar(file.stan, file.info(file.stan)$size)
stan.mod <- stan( model_code = stan.code, data = dat, chains = 4,
                  iter = 3500, warmup = 500, seed = 779 )
```

```
> print(stan.mod, pars = c("alpha", "sigma2b1", "sigma2b"), digits = 3)
```

4 chains, each with iter=3500; warmup=500; thin=1;
 post-warmup draws per chain=3000, total post-warmup draws=12000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha[1]	1.766	0.002	0.114	1.540	1.692	1.767	1.842	1.985	5557	1.000
alpha[2]	0.880	0.002	0.137	0.612	0.788	0.881	0.971	1.149	5020	1.000
alpha[3]	-0.958	0.006	0.420	-1.790	-1.239	-0.958	-0.678	-0.120	5144	1.000
alpha[4]	0.353	0.003	0.215	-0.069	0.212	0.351	0.495	0.777	4489	1.000
alpha[5]	0.479	0.005	0.371	-0.252	0.237	0.483	0.723	1.206	6573	1.001
alpha[6]	-0.103	0.001	0.087	-0.275	-0.161	-0.103	-0.045	0.067	11302	1.000
sigma2b1	0.253	0.001	0.072	0.139	0.201	0.244	0.294	0.419	6698	1.001
sigma2b	0.135	0.001	0.032	0.081	0.112	0.132	0.155	0.206	1599	1.001

The fixed effects in the output above correspond to the following:

- ▶ α_1 : intercept,
- ▶ α_2 : $\log(\text{baseline}/4)$ (centered),
- ▶ α_3 : treatment,
- ▶ α_4 : treatment-by- $\log(\text{baseline}/4)$ interaction,
- ▶ α_5 : $\log(\text{age})$ (centered),
- ▶ α_6 : V4.

In SAS, we use the BGLIMM procedure:

```
* Read in data;
data epil;
    infile "epilepsy.dat" firstobs=2;
    input y trt base age V4 subject period;
run;

* Create new variable log(base/4), trt-by-log(base/4) interaction, and subject-by-visit indicator;
data epil2;
    set epil;
    lbase4 = log(base/4);
    trt_lbase4 = trt * lbase4;
    lage = log(age);
    sub_by_visit = _n_;
run;

* Center lbase4 and lage covariates;
proc standard data=epil2 mean=0 out=epil3;
    var lbase4 lage;
run;

* Fit mixed model with BGLIMM;
proc bglimm data=epil3 seed=779;
    class subject sub_by_visit;
    model y = lbase4 trt trt_lbase4 lage V4 / dist=poisson link=log cprior=normal(var=1e4);
    random int / sub = subject covprior=igamma(shape=.001, scale=.001);
    random int / sub = sub_by_visit covprior=igamma(shape=.001, scale=.001);
run;
```

The results for the model using BGLIMM with 5000 iterations and a burn-in of 500 are included in the table below.

Parameter	Mean	Standard Deviation	95% HPD Interval	
Intercept	1.7703	0.1172	1.5534	2.0229
log(Baseline/4) (centered)	0.8728	0.1419	0.5974	1.1595
Treatment	-0.9842	0.4006	-1.7728	-0.1896
Treatment × log(Baseline/4)	0.3642	0.1965	-0.0302	0.7327
log(Age) (centered)	0.4862	0.3680	-0.2268	1.2249
V4	-0.0982	0.0886	-0.2773	0.0659
σ_{b1}^2	0.2486	0.0730	0.1242	0.3947
σ_b^2	0.1371	0.0321	0.0787	0.2025

Comparing the output from both models, we see that `rstan` and BGLIMM provide similar posterior results.

Chapter 6:

Nonparametric and Semi-Parametric Bayesian Methods

Nonparametric Bayesian Methods

The general problem of nonparametric inference can be stated as follows. Suppose x_1, \dots, x_n is a sample from $f(\cdot)$, where $f(\cdot)$ is an unknown density function of the observations. We wish to make inferences about $f(\cdot)$, such as estimate $f(\cdot)$. The nonparametric density estimate of $f(\cdot)$ takes the form

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - x_i}{h}\right)$$

where h is a smoothing parameter called the *bandwidth*, and K is called the *kernel*. We refer to $\hat{f}(t)$ as the **kernel density estimator**.

There are also methods for nonparametric estimation of a cumulative distribution function $F(\cdot)$.

The basic **building tool** for Bayesian nonparametric methods is called the **Dirichlet process**. The Dirichlet process provides the Bayesian with a nonparametric prior specification over the class of possible distribution functions $F(x)$ for a random variable X , where $F(x) = P(X \leq x)$.

To discuss the Dirichlet process, we first need to discuss the **Dirichlet distribution**. The Dirichlet distribution is the multivariate generalization of the beta distribution. Let z_1, \dots, z_k be independent random variables, with

$$z_j \sim \text{gamma}(\alpha_j, 1), \quad j = 1, \dots, k.$$

Define

$$u = \sum_{j=1}^k z_j ,$$

and

$$y_j = \frac{z_j}{u} = \frac{z_j}{\sum_{j=1}^k z_j}.$$

Since

$$\sum_{j=1}^k y_j = 1,$$

the distribution of y_1, \dots, y_k is **singular**. Thus it is usually more convenient to work with the non-singular distribution of (y_1, \dots, y_{k-1}) .

Definition 6.1

(y_1, \dots, y_{k-1}) defined above have a **k-1 dimensional Dirichlet distribution**, denoted by $D_{k-1}(\alpha_1, \dots, \alpha_k)$.

To obtain the joint density of (y_1, \dots, y_{k-1}) , note that the transformation from z_1, \dots, z_k into (u, y_1, \dots, y_{k-1}) is one-to-one. This is the region $z_j > 0$ for all j mapping into $u > 0$, $y_j > 0$, $j = 1, \dots, k-1$, and $\sum_{j=1}^{k-1} y_j \leq 1$.

The inverse transformation is

$$z_j = uy_j, \quad j = 1, \dots, k-1,$$

$$z_k = u(1 - \sum_{j=1}^{k-1} y_j).$$

The Jacobian of the transformation is

$$\left| \frac{\partial z}{\partial(y, u)} \right| = u^{k-1}.$$

This leads to the joint density

$$f(y_1, \dots, y_{k-1}, u) = \left(\frac{1}{c} \right) (u^{\alpha-1} \exp\{-u\}) \left(\prod_{j=1}^{k-1} y^{\alpha_j-1} \right) \left(1 - \sum_{j=1}^{k-1} y_j \right)^{\alpha_{k-1}},$$

over the support region mentioned above.

We see that $u \sim \text{gamma}(\alpha, 1)$, where $\alpha = \sum_{j=1}^k \alpha_j$, and u is **independent** of y_1, \dots, y_{k-1} . The constant c is the normalizing constant.

The normalizing constant for y_1, \dots, y_{k-1} is available from the z_j and u normalizing constants, and the density of $(y_1, \dots, y_{k-1}) \sim D_{k-1}(\alpha_1, \dots, \alpha_k)$ is

$$f(y_1, \dots, y_{k-1}) = \left(\frac{\Gamma(\alpha)}{\prod_{j=1}^k \Gamma(\alpha_j)} \right) \left(\prod_{j=1}^{k-1} y_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} y_j \right)^{\alpha_k - 1}$$

with support over the $k - 1$ **dimensional simplex**,

$$y_j \geq 0, \quad j = 1, \dots, k - 1, \quad \sum_{j=1}^{k-1} y_j \leq 1.$$

Formally, we have the following definition.

Definition 6.2

$(y_1, \dots, y_{k-1}) \sim D_{k-1}(\alpha_1, \dots, \alpha_k)$ if

$$f(y_1, \dots, y_{k-1}) = \left(\frac{\Gamma(\alpha)}{\prod_{j=1}^k \Gamma(\alpha_j)} \right) \left(\prod_{j=1}^{k-1} y_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} y_j \right)^{\alpha_k - 1}$$

where

$$y_j \geq 0, \quad j = 1, \dots, k-1, \quad \sum_{j=1}^{k-1} y_j \leq 1, \quad \alpha = \sum_{j=1}^k \alpha_j.$$

The joint density of (y_1, \dots, y_{k-1}) above is called the $k-1$ dimensional Dirichlet density.

The **Dirichlet distribution** is the **conjugate prior** distribution for the **multinomial** family. That is, if $(x_1, \dots, x_k) \sim \text{Multinomial}(p_1, \dots, p_k)$,

where $\sum_{j=1}^k p_j = 1$, then the conjugate prior for (p_1, \dots, p_{k-1}) is

$$(p_1, \dots, p_{k-1}) \sim D_{k-1}(\alpha_1, \dots, \alpha_k).$$

Thus a posteriori,

$$(p_1, \dots, p_{k-1} \mid x) \sim D_{k-1}(\alpha_1 + x_1, \dots, \alpha_k + x_k),$$

where $x = (x_1, \dots, x_k)$.

More formally, this can be stated in the following theorem.

Theorem 6.1

The Dirichlet distribution is the conjugate prior distribution for the multinomial family.

Proof:

Suppose $(x_1, \dots, x_k) \sim \text{Multinomial}(p_1, \dots, p_k)$ where $\sum_{j=1}^k p_j = 1$. Then (x_1, \dots, x_k) have a $k - 1$ dimensional multinomial distribution, with likelihood function

$$L(p_1, \dots, p_{k-1}) = \left(\prod_{j=1}^{k-1} p_j^{x_j} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{x_k}.$$

Suppose we specify a Dirichlet prior for (p_1, \dots, p_{k-1}) , that is we take $(p_1, \dots, p_{k-1}) \sim D_{k-1}(\alpha_1, \dots, \alpha_k)$.

Then

$$\begin{aligned}
 p(p_1, \dots, p_k \mid x) &\propto L(p_1, \dots, p_{k-1})\pi(p_1, \dots, p_{k-1}) \\
 &\propto \left(\prod_{j=1}^{k-1} p_j^{x_j} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{x_k} \left(\prod_{j=1}^{k-1} p_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{\alpha_k - 1} \\
 &= \left(\prod_{j=1}^{k-1} p_j^{x_j + \alpha_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{x_k + \alpha_k - 1} \\
 &\propto D_{k-1}(\alpha_1 + x_1, \dots, \alpha_k + x_k).
 \end{aligned}$$

Thus

$$(p_1, \dots, p_{k-1} \mid x) \sim D_{k-1}(\alpha_1 + x_1, \dots, \alpha_k + x_k).$$

Properties of the Dirichlet Distribution

- 1) Univariate marginal distributions are beta distributions. It follows that $y_j \sim \text{beta}(\alpha_j, \alpha - \alpha_j)$, $j = 1, \dots, k - 1$. This can also be written

$$y_j \sim D_1(\alpha_j, \alpha - \alpha_j).$$

- 2) Bivariate distributions: by a similar argument, we can show that $(y_j, y_k) \sim D_2(\alpha_j, \alpha_k, \alpha - \alpha_j - \alpha_k)$. In general, **all marginal distributions of Dirichlet are Dirichlet**.
- 3) **Pooling**: using properties of the gamma distribution, it follows that, for example,

$$(y_1 + y_2, y_3 + y_4 + y_5) \sim D_2(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4 + \alpha_5, \alpha - \sum_{j=1}^5 \alpha_j).$$

- 4) **Moments:** $E(y_j)$, $E(y_j^2)$, etc... are available from the beta distribution. It is also easily shown that

$$E(y_j) = \frac{\alpha_j}{\alpha}, \quad E(y_j^2) = \frac{\alpha_j(\alpha_j + 1)}{\alpha(\alpha + 1)}, \quad E(y_i y_j) = \frac{\alpha_i \alpha_j}{\alpha(\alpha + 1)}$$
$$\text{Var}(y_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}, \quad \text{Cov}(y_i, y_j) = -\frac{\alpha_i \alpha_j}{\alpha^2(\alpha + 1)}.$$

In Bayesian nonparametric inference, the typical approach is to specify a prior distribution over the space of **all possible cumulative distribution functions**, $F(x)$, $0 \leq F(x) \leq 1$. The prior that is most popular is the **Dirichlet process prior**.

Ferguson (1973, 1974, *Annals of Statistics*) introduced the **Dirichlet process**.

Let the sample space be denoted by S , and suppose $S = B_1 \cup \dots \cup B_k$, where the B_j 's are disjoint. The B_j 's, for example, can be disjoint intervals. Then a stochastic process P indexed by elements of a particular partition $B = \{B_1, \dots, B_k\}$ is said to be a Dirichlet process on (S, B) with parameter vector α , if for **any** partition of S , $S = B \cup \dots \cup B_k$, the random vector $(P(B_1), \dots, P(B_k))$ has a **Dirichlet distribution** with parameter $(\alpha(B_1), \dots, \alpha(B_k))$.

We can make the definition a bit less general and write for any partition $S = B_1 \cup \dots \cup B_k$, $(F(B_1), \dots, F(B_k))$ has a Dirichlet distribution with parameters $(\alpha(B_1), \dots, \alpha(B_k))$. The B_j 's for example might be disjoint intervals of the form $B_j = (b_{1j} b_{2j}]$, so that $F(B_j) = F(b_{2j}) - F(b_{1j})$ and $\alpha(B_j) = \alpha(b_{2j}) - \alpha(b_{1j})$.

The parameter vector α is a **probability measure**, i.e., a distribution function itself so that we can write $\alpha = F_0(\cdot)$, where $F_0(\cdot)$ is the prior hyperparameter for $F(\cdot)$. Thus $\alpha(B_j) = F_0(b_{2j}) - F_0(b_{1j})$.

We can define a weight parameter c_0 ($c_0 > 0$) that gives prior weight to $F_0(\cdot)$, so that we can write $(F(B_1), \dots, F(B_k))$ has a Dirichlet distribution with parameters $(c_0 F_0(B_1), \dots, c_0 F_0(B_k))$.

Finally, we say that F has a Dirichlet process prior with parameter $c_0 F_0$ if $(F(B_1), \dots, F(B_k))$ has a Dirichlet distribution with parameters $(c_0 F_0(B_1), \dots, c_0 F_0(B_k))$ for **every possible partition** of the sample space $S = B_1 \cup \dots \cup B_k$. We denote the **Dirichlet process prior** by DP.

Theorem 6.2

Suppose x_1, \dots, x_n is a sample from $F(\cdot)$, and suppose a priori $F \sim DP(c_0 F_0)$. Then

$$F \mid x \sim DP\left(c_0 F_0 + \sum_{i=1}^n \delta_{x_i}\right),$$

where δ_{x_i} is a **point mass** giving probability 1 to x_i .

Thus, a posteriori, the distribution of F is a Dirichlet process with parameter

$$c_0 F_0 + \sum_{i=1}^n \delta_{x_i} = c_0 F_0 + n F_n,$$

where F_n is the empirical c.d.f.

We note that $\sum_{i=1}^n \delta_{x_i} = n F_n$, where $F_n(\cdot)$ is the **empirical c.d.f.**

$$F_n(x) = \frac{\# \text{ of } x_i \text{ in the sample } \leq x}{n}.$$

In an actual data analysis, we construct the intervals B_1, \dots, B_k after looking at the data x_1, \dots, x_n , and construct them in such a way that B_j has at least one of the x_i 's.

Example 6.1

Suppose x_1, \dots, x_5 are i.i.d. from $F(\cdot)$, and suppose $F \sim DP(c_0 F_0)$ where $c_0 = .1$ and $F_0 = \text{exponential}(1)$, i.e., $F_0(x) = 1 - \exp\{-x\}$.

Further, suppose $x_1 = 1, x_2 = .7, x_3 = .8, x_4 = 1.2, x_5 = 1.3$, and let

$$\begin{aligned}B_1 &= \{x : 0 < x \leq 1\} \\B_2 &= \{x : 1 < x \leq 1.25\} \\B_3 &= \{x : 1.25 < x < \infty\}.\end{aligned}$$

We note that the intervals are always left open and right closed. Thus,

$$\begin{aligned} F(B_1) &= F(1) - F(0) = p_1 \\ F(B_2) &= F(1.25) - F(1) = p_2 \\ F(B_3) &= F(\infty) - F(1.25) = p_3, \end{aligned}$$

where $p_1 + p_2 + p_3 = 1$.

Therefore,

$$(p_i, p_j) \sim \text{Dirichlet}(c_0 F_0(B_1), c_0 F_0(B_2), c_0 F_0(B_3)),$$

where

$$\begin{aligned} F_0(B_1) &= F_0(1) - F_0(0) = 1 - \exp\{-1\} = .632 \\ F_0(B_2) &= F_0(1.25) - F_0(1) = \exp\{-1\} - \exp\{-1.25\} = .081 \\ F_0(B_3) &= F_0(\infty) - F_0(1.25) = \exp\{-1.25\} = .287. \end{aligned}$$

Thus, a priori

$$(p_1, p_2) \sim D_2(.0632, .0081, .0287).$$

By the theorem,

$$(p_1, p_2) \mid x \sim D_2(c_0 F_0(B_1) + n F_n(B_1), c_0 F_0(B_2) + n F_n(B_2), c_0 F_0(B_3) + n F_n(B_3)).$$

A different partition of the sample space leads to a different definition of the p_j 's and hence a Dirichlet prior with different prior parameters. Now

$$\begin{aligned} F_n(x) &= \frac{\# \text{ of } x_j \leq x}{5} \\ F_n(B_1) &= F_n(1) - F_n(0) = \frac{3}{5} - 0 = \frac{3}{5} \\ F_n(B_2) &= F_n(1.25) - F_n(1) = \frac{4}{5} - \frac{3}{5} = \frac{1}{5} \\ F_n(B_3) &= F_n(\infty) - F_n(1.25) = 1 - \frac{4}{5} = \frac{1}{5}, \end{aligned}$$

where $F_n(B_1) + F_n(B_2) + F_n(B_3) = 1$.

The distribution of $F_n(x)$ is shown in Figure 6.1.

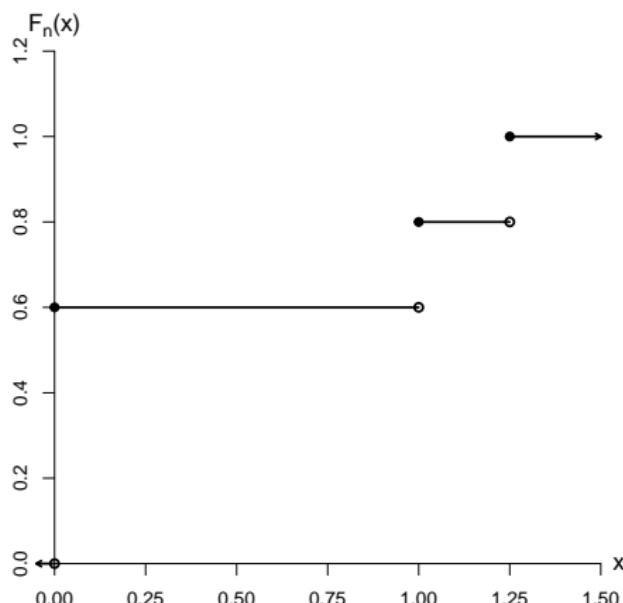


Figure 6.1:

Thus,

$$\begin{aligned}(p_1, p_2) \quad | \quad x &\sim D_2 \left(.0632 + 5 \left(\frac{3}{5} \right), .0081 + 5 \left(\frac{1}{5} \right), .0287 + 5 \left(\frac{1}{5} \right) \right) \\ &= D_2(3.0632, 1.0081, 1.0287).\end{aligned}$$

The parameter $F_0(\cdot)$ is often called the **base measure** of the **Dirichlet process prior**, and the prior parameter c_0 is called the **weight parameter** of the **base measure**.

One can use the DP prior on the c.d.f. to construct nonparametric Bayesian density estimates of $f(x)$, the density corresponding to $F(x)$. See (Hjort, 1996, *Bayesian Statistics 5*).

Also, the Dirichlet process prior can be used in Bioassay problems in which $x_i \sim \text{Bin}(n_i, p_i)$, $p_i = F(t_i)$, t_i is the dose level for individual i , and $x_i = \#$ of subjects with positive response at dose t_i . Here, $F(\cdot)$ is typically unknown and we can take $F(\cdot)$ to have a DP prior. For more on this type of problem, see Gelfand and Kuo (1991, *Biometrika*).

The Bayesian nonparametric literature has grown in the past decade. Notable articles include Susarla and Van Ryzin (1976, *JASA*), Kalbfleisch (1977, *JRSS-B*), Dykstra and Laud (1981, *Annals of Statistics*), Bush and MacEachern (1996, *Biometrika*), Doss (1994, *Annals of Statistics*), Escobar (1994, *JASA*), Escobar and West (1995, *JASA*), Kleinman and Ibrahim (1998, *Biometrics*) and Kleinman and Ibrahim (1998, *Statistics in Medicine*).

Semi-Parametric Bayesian Inference

In many parametric likelihood models, we often wish to relax the assumption of a parametric prior on the parameters. In these cases, it may be useful to consider a non-parametric prior on the parameters, such as a Dirichlet process prior, which leads to **mixtures of Dirichlet processes**. Before turning to the semi-parametric model, first consider the fully parametric situation. Suppose y_i is an $n_i \times 1$ random vector indexed by the $p \times 1$ parameter vector θ_i , for each $i = 1, \dots, n$.

Suppose the θ_i have a prior distribution with hyperparameter θ_0 . That is, $\theta_i \stackrel{\text{i.i.d.}}{\sim} G(\cdot | \theta_0)$. If $G(\cdot | \theta_0)$ is a **specified** function, then this corresponds to the fully parametric situation. The fully parametric situation can be described by two stages:

Stage 1: $[y_i | \theta_i]$ (parametric likelihood function)
 $(i = 1, \dots, n)$

Stage 2: $[\theta_i | \theta_0] = G(\cdot | \theta_0)$,

where $G(\cdot | \theta_0)$ is a **specified** prior distribution. That is, $G(\cdot | \theta_0)$ is a known functional form, such as a normal, gamma, exponential, beta, etc.

The **Mixtures of Dirichlet Process** (MDP) model removes the parametric assumption on $G(\cdot | \theta_0)$, that is $G(\cdot | \theta_0)$ is **not known**, and thus no functional form is specified for G . Thus the MDP model has 3 stages

Stage 1: $[y_i | \theta_i]$ (parametric likelihood)

Stage 2: $\theta_i \stackrel{\text{i.i.d.}}{\sim} G$ (G unknown)

Stage 3: $G | c_0, G_0 \sim DP(c_0 | G_0)$,

where $c_0 > 0$, and G_0 is the base measure. Note that G_0 is a parametric function which will depend on its own known parameters. For example, G_0 can be,

- ▶ $G_0 = N(\mu_0, \sigma_0^2)$
- ▶ $G_0 = \text{exponential}(\lambda_0)$

and so on. Thus the MDP model has 3 stages with the last stage being the DP specification. The **base measure** G_0 is a distribution which approximates the true non-parametric shape of G . The scalar c_0 reflects our prior belief about how similar the non-parametric distribution G is to the base measure G_0 .

There are two special cases, in which the MDP model leads to the fully parametric case. As $c_0 \rightarrow \infty$, $G \rightarrow G_0$, so that the base measure is the (parametric) prior distribution for θ_i . Also if $\theta_i \equiv \theta$ for all i , the same is also true, i.e., $G = G_0$.

Thus MDP models are useful only in models in which we have a **parameter for every observation**, i.e., for likelihoods of the form $[y_i | \theta_i]$. For a mere hierarchical modeling approach, it is possible to place prior distributions on (c_0, θ_0) , where θ_0 is the hyperparameter for G_0 . Thus the specification given above is semi-parametric in the sense that a parametric likelihood specification is given in stage 1, and a non-parametric specification is given in stages 2 and 3.

The Posterior Distribution of the θ_i 's

The conditional posterior distribution of the θ_i 's is a mixture distribution, and can be derived by the **Polya urn representation** of the Dirichlet process. See MacEachern (1994, *Comm. Stat., Comp. and Sim.*) and Escobar (1994, *JASA*). The Polya urn representation of the Dirichlet process is quite useful for sampling purposes.

It is described as follows:

- (1) The draw of θ_1 is always from the base measure G_0 .
- (2) The draw of θ_2 is equal to θ_1 with probability p_1 and is from the base measure with probability $p_0 = 1 - p_1$. The draw of θ_3 is equal to θ_1 with probability p_1 , θ_2 with probability p_2 , and is a draw from the base measure with probability $p_0 = 1 - p_1 - p_2$. The values of the p_i 's change with every new draw. The process continues until θ_n is equal to each of the preceding θ_i 's with probability p_i , $i = 1, \dots, n-1$, and is a draw from the base measure with probability $p_0 = 1 - \sum_{i=1}^{n-1} p_i$.

The values of the p_i 's, $i = 0, \dots, n-1$, are determined from the Dirichlet process parameters. In other words, the θ_i 's are actually drawn from a mixture distribution where the mixing probabilities are determined by the Dirichlet process of Stage 3, thus giving rise to the MDP label.

From this representation, it is clear that if all of the $\theta_i = \theta$ for all i , then we **draw θ from the base measure with probability 1**, and thus the base measure **is** the prior.

The MDP model is simplified in practice by the Polya urn representation, using the fact that marginally, the θ_i 's are distributed as the base measure along with the added property that $P(\theta_i = \theta_j) > 0$ for $i \neq j$. The Dirichlet process prior results in what MacEachern (1994) calls a **cluster structure** among the θ_i 's. This cluster structure partitions the $n \theta_i$'s into k sets or **clusters**, $0 < k \leq n$. All of the observations in a cluster share an identical value of θ and subjects in different clusters have different values of θ .

As described by Escobar (1994, *JASA*), conditional on the other θ_j 's, θ_i has the following **mixture distribution**:

$$p(\theta_i | y, \theta_j, j \neq i) \propto \sum_{j \neq i} q_j \delta_{\theta_j} + c_0 q_0 g_0(\theta_i) p(y_i | \theta_i),$$

where $y = (y_1, \dots, y_n)$, and $p(y_i | \theta_i)$ is the sampling density of the θ_i 's. The values q_j and $c_0 q_0$ can be normalized to get the selection probabilities p_j , $j = 0, \dots, n - 1$, in the urn scheme described above.

Also, δ_{θ_j} is a **point mass** at θ_j , and $g_0(\cdot)$ is the density corresponding to $G_0(\cdot)$. Additionally,

$$\begin{aligned} q_j &= p(y_i \mid \theta_j) , \quad j = 1, \dots, i-1, i+1, \dots n, \\ q_0 &= \int p(y_i \mid \theta_i) g_0(\theta_i) d\theta_i . \end{aligned}$$

From the previous pages, we have

$$\begin{aligned} p(\theta_i | y, \theta_j, i \neq j) &\propto \left\{ \sum_{\substack{j \neq i \\ j=1}}^n q_j \delta_{\theta_j} \right\} + c_0 q_0 g_0(\theta_i) p(y_i | \theta_i) \\ &= \sum_{\substack{j \neq i \\ j=1}}^n q_j \delta_{\theta_j} + c_0 q_0^2 \left[\frac{g_0(\theta_i) p(y_i | \theta_i)}{q_0} \right] \\ &= \sum_{\substack{j \neq i \\ j=1}}^n q_j \delta_{\theta_j} + c_0 q_0^2 p(\theta_i | y_i) , \end{aligned} \tag{6.1}$$

where

$$\begin{aligned} q_0 &= \int g_0(\theta_i) p(y_i|\theta_i) d\theta_i \\ q_j &= p(y_i|\theta_j), \quad j = 1, \dots, n, \quad j \neq i. \end{aligned}$$

We see that (6.1) consists of a mixture distribution involving n components: $n - 1$ of the components are point masses at θ_j , and the last component of the mixture is a continuous density $p(\theta|y_i)$ where

$$p(\theta_i|y_i) = \frac{p(y_i|\theta_i)g_0(\theta_i)}{\int p(y_i|\theta_i)g_0(\theta_i)} = \text{posterior distribution of } \theta_i.$$

Let us now normalize the selection probabilities.

The normalized q_j 's can be written as

$$p_j = \frac{q_j}{c_0 q_0^2 + \sum_{\substack{r=1 \\ r \neq i}}^n q_r}, \quad j = 1 \dots n, \quad j \neq i.$$

The normalized $c_0 q_0^2$ is

$$p_0 = \frac{c_0 q_0^2}{c_0 q_0^2 + \sum_{\substack{r=1 \\ r \neq i}}^n q_r} \equiv 1 - \sum_{\substack{j=1 \\ j \neq i}}^n p_j .$$

Notice that

$$p_0 + \sum_{\substack{j=1 \\ j \neq i}}^n p_j = 1$$

since

$$p_0 + \sum_{\substack{j=1 \\ j \neq i}}^n p_j = \frac{c_0 q_0^2}{c_0 q_0^2 + \sum_{\substack{r=1 \\ r \neq i}}^n q_r} + \frac{\sum_{\substack{j=1 \\ j \neq i}}^n q_j}{c_0 q_0^2 + \sum_{\substack{r=1 \\ r \neq i}}^n q_r} = 1 .$$

Now the normalized mixture distributions are written as

$$\begin{aligned}
 p(\theta_i|y, \theta_j, i \neq j) &= \left\{ \sum_{\substack{j=1 \\ j \neq i}}^n p_j \delta_{\theta_j} \right\} + p_0 p(\theta_i|y_i) \\
 p_j &= \frac{q_j}{c_0 q_0^2 + \sum_{\substack{r=1 \\ r \neq i}}^n q_r}, \quad j = 1, \dots, n, j \neq i \\
 p_0 &= 1 - \sum_{\substack{j=1 \\ j \neq i}}^n p_j = \frac{c_0 q_0^2}{c_0 q_0^2 + \sum_{\substack{r=1 \\ r \neq i}}^n q_r}.
 \end{aligned} \tag{6.2}$$

Another way of writing (6.2) is

$$\theta_i = \begin{cases} \theta_j & \text{w.p. } p_j, \quad j = 1, \dots, n, j \neq i \\ \text{a sample from } [\theta_i|y_i] & \text{w.p. } p_0 \end{cases}.$$

For example, if

$$\begin{aligned}[y_i|\theta_i] &= N(\theta_i, \sigma_y^2) \\ \theta_i &\sim G \\ G &\sim DP(c_0 G_0) \\ G_0 &= N(\mu, \sigma_\theta^2),\end{aligned}$$

then we have

$$\begin{aligned}[\theta_i|y_i] &= N(a, b) \\ a &= (\sigma_\theta^2 + \sigma_y^2)^{-1} \sigma_\theta^2 \sigma_y^2 \left(\frac{\mu}{\sigma_\theta^2} + \frac{y_i}{\sigma_y^2} \right) \\ b &= (\sigma_\theta^2 + \sigma_y^2)^{-1} \sigma_\theta^2 \sigma_y^2\end{aligned}$$

so that $p(\theta_i|y_i) = (2\pi)^{-1/2} b^{-1/2} \exp \left\{ -\frac{1}{2b} (\theta_i - a)^2 \right\}$.

To demonstrate the MDP model, consider the following simple example. Suppose $y_i | \theta_i \sim N(\theta_i, \sigma_y^2)$, where σ_y^2 is known. In this case $n_i = 1$, $i = 1, \dots, n$. Also assume each θ_i has the univariate normal distribution (fully parametric specification), so that

$$\begin{aligned}\text{Stage 1: } [y_i | \theta_i, \sigma_y^2] &= N(\theta_i, \sigma_y^2) \\ \text{Stage 2: } [\theta_i | \mu, \sigma_\theta^2] &= N(\mu, \sigma_\theta^2).\end{aligned}$$

The MDP model **removes** the assumption of normality at the second stage, resulting in

$$\begin{aligned}\text{Stage 1: } [y_i | \theta_i, \sigma_y^2] &= N(\theta_i, \sigma_y^2) \\ \text{Stage 2: } \theta_i &\sim G \quad (\text{i.i.d.}) \\ \text{Stage 3: } [G | c_0, G_0] &\sim DP(c_0 | G_0).\end{aligned}$$

Typical G_0 's in this case would be $G_0 = N(\mu, \sigma_\theta^2)$, so that the base measure hyperparameters are (μ, σ_θ^2) .

With $G_0 = N(\mu, \sigma_\theta^2)$, the unnormalized selection probability is $q_j = p(y_i | \theta_j) \equiv \phi(y_i | \theta_j, \sigma_y^2)$ where $\phi(\cdot | \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2 .

With probability proportional to q_j , $\theta_i \sim \delta_{\theta_j}$, which means that $\theta_i = \theta_j$ with probability 1.

The unnormalized selection probability q_0 is given by

$$\begin{aligned} q_0 &= \int p(y_i | \theta_i, \sigma_y^2) g_0(\theta_i | \theta_0) d\theta_i \\ &= \int \phi(y_i | \theta_i, \sigma_y^2) \phi(\theta_i | \mu, \sigma_\theta^2) d\theta_i. \end{aligned}$$

With probability proportional to $c_0 q_0$,

$$\begin{aligned} [\theta_i | y_i] &\sim g_0(\theta_i) p(y_i | \theta_i) \\ &= N(\theta_i | \mu, \sigma_\theta^2) N(y_i | \theta_i, \sigma_y^2), \end{aligned}$$

and thus, $[\theta_i | y_i] = N(a, b)$, where

$$\begin{aligned} a &= (\sigma_\theta^2 + \sigma_y^2)^{-1} \sigma_\theta^2 \sigma_y^2 \left(\frac{\mu}{\sigma_\theta^2} + \frac{y_i}{\sigma_y^2} \right), \\ b &= (\sigma_\theta^2 + \sigma_y^2)^{-1} \sigma_\theta^2 \sigma_y^2. \end{aligned}$$

In this previous example, selecting G_0 to be normal emulates the conjugate relationship between the sampling distribution and prior in the usual Bayesian hierarchy. In the MDP case, the sampling distribution is **conjugate** to the **base measure**. MDP models with base measures and sampling distributions that are conjugate in this fashion are called **conjugate MDP models**.

The computational advantages of conjugate MDP models are clear from the example above. First, q_0 has a closed form. Second, the posterior distribution of θ_i corresponding to q_0 is from the same exponential family as the base measure. As a result, Gibbs sampling in the conjugate model can proceed in a straightforward fashion.

DP Priors in the Normal Linear Random Effects Model

Consider the usual normal linear random effects model

$$y_i = X_i\beta + Z_i b_i + \epsilon_i ,$$

where $\epsilon_i \sim N_{n_i}(0, \sigma^2 I_{n_i})$, $i = 1, \dots, n$. Let $\tau = 1/\sigma^2$. The MDP model is well suited for the random effects model since the b_i 's are different for each observation.

The MDP model then proceeds as follows:

$$\begin{aligned} [y_i \mid \beta, b_i, y] &\sim N_{n_i}(X_i\beta + Z_i b_i, \tau^{-1} I_{n_i}) \\ \tau &\sim \text{gamma}\left(\frac{\alpha_0}{2}, \frac{\lambda_0}{2}\right) \\ \beta &\sim N_p(\mu_0, \Sigma_0) \\ b_i &\sim G \quad (\text{i.i.d.}) \quad (b_i \text{ is } v \times 1) \\ G &\sim DP(c_0, G_0), \end{aligned}$$

where $G_0 = N_v(0, D)$.

We can sample from this model using the Gibbs sampler. The full conditionals of β and τ are given by

$$[\beta \mid b, \tau, y] \sim N_p(Ta, T) ,$$

where

$$T = \left(\tau \sum_{i=1}^n X_i' X_i + \Sigma_0^{-1} \right)^{-1} ,$$

$$a = \tau \sum_{i=1}^n X_i'(y_i - Z_i b_i) + \Sigma_0^{-1} \mu_0 ,$$

and

$$[\tau \mid \beta, b, y] \sim \text{gamma} \left(\frac{N + \alpha_0}{2}, \frac{\lambda_0 + \sum_{i=1}^n r_i' r_i}{2} \right) ,$$

where $r_i = y_i - X_i \beta - Z_i b_i$.

These full conditionals are unchanged by the MDP model. We can get the conditional posterior distributions of b_i as

$$\begin{aligned} p(b_i | \beta, \tau, y, b_j, i \neq j) &\propto \sum_{j \neq i} \phi(y_i | X_i \beta + Z_i b_j, \tau^{-1} I_{n_i}) \delta_{b_j} \\ &\quad + \left\{ c_0 \int \phi(y_i | X_i \beta + Z_i b_i, \tau^{-1} I_{n_i}) \phi(b_i | 0, D) db_i \right\} \\ &\quad \times \phi(b_i | 0, D) p(y_i | b_i, \beta, \tau). \end{aligned}$$

Thus, we are led to

$$\begin{aligned} p(b_i | \beta, \tau, y, b_j, j \neq i) &\propto \left(\sum_{j \neq i} \tau^{\frac{n_i}{2}} \exp \left\{ -\frac{\tau}{2} (y_i - X_i \beta - Z_i b_j)' (y_i - X_i \beta - Z_i b_j) \right\} \delta_{b_j} \right) \\ &\quad + c_0 |Q_i|^{\frac{1}{2}} |D|^{-\frac{1}{2}} \tau^{\frac{n_i}{2}} \exp \left\{ \frac{\tau}{2} (y_i - X_i \beta)' U_i (y_i - X_i \beta) \right\} \\ &\quad \times \phi(b_i | 0, D) p(y_i | b_i, \beta, \tau), \end{aligned}$$

where $Q_i = (D^{-1} + \tau Z_i' Z_i)^{-1}$ and $U_i = \tau Z_i Q_i Z_i' - I$.

In the specification above, each summand is separated into 2 elements. The first element is a mixing probability, and the second is a distribution to be mixed.

So with probability proportional to

$$\tau^{\frac{n_i}{2}} \exp \left\{ -\frac{\tau}{2} (y_i - X_i \beta - Z_i b_j)'(y_i - X_i \beta - Z_i b_j) \right\} ,$$

we select from the distribution δ_{b_j} , which means that we set $b_i = b_j$ with probability 1.

Also, with probability proportional to

$$c_0 | Q_i |^{\frac{1}{2}} | D |^{-\frac{1}{2}} \tau^{\frac{n_i}{2}} \exp \left\{ \frac{\tau}{2} (y_i - X_i \beta)' U_i (y_i - X_i \beta) \right\} ,$$

we select from

$$p(b_i | \beta, \tau, y_i) \propto \phi(b_i | 0, D) p(y_i | b_i, \beta, \tau) ,$$

which means that we sample b_i from its full conditional

$$[b_i | \beta, \tau, y_i] \sim N_v(\tau Q_i Z'_i (y_i - X_i \beta), Q_i) .$$

This results in a mixture distribution where one piece is a normal distribution and all of the others are point masses. The Gibbs sampler for $p(\beta, b, \tau | D, y)$ can now be described as follows:

0. Select starting values $b^{(0)}$ and $\tau^{(0)}$. Set $i = 0$.
1. Sample $\beta^{(i+1)}$ from $[\beta | b^{(i)}, \tau^{(i)}, y]$ according to $[\beta | b, \tau, y] \sim N_p(Ta, T)$.
2. Sample $\tau^{(i+1)}$ from $[\tau | \beta^{(i+1)}, b^{(i)}, y]$
according to $[\tau | \beta, b, y] \sim \text{gamma} \left(\frac{n+\alpha_0}{2}, \frac{\lambda_0 + \sum_{i=1}^n r'_i r_i}{2} \right)$.
- 3.1 Sample $b_1^{(i+1)}$ from $[b_1 | \beta^{(i+1)}, b_j^{(i)}, j \neq 1, y]$ according to the mixture distribution $p(b_i | \beta, \tau, y, b_j, j \neq i)$ given earlier.
- ⋮
- ⋮
- 3.n Sample $b_n^{(i+1)}$ from $[b_n | \beta^{(i+1)}, b_j^{(i+1)}, j \neq n, \tau^{(i+1)}, y]$.
4. Set $i = i + 1$ and return to step 1.

Since D^{-1} is typically unknown, we need to specify a prior for it. We take

$$D^{-1} \sim W_v(\nu_0, C_0).$$

After sampling the random effects for each subject, the subjects will be grouped into **clusters** in which the subjects have equal b_i 's. That is, after selecting a new b_i for each subject i in the sample, there will be some number k , $0 < k \leq n$ of unique values among the b_i 's. Denote these unique values by γ_l , $l = 1, \dots, k$. Additionally, let l represent the set of subjects with common random effect γ_l .

Note that knowing the random effects is equivalent to knowing k , all of the γ_l 's, and the cluster memberships l . Then for the purposes of calculating the full conditionals of D^{-1} , the γ_l are k independent observations from $N_v(0, D)$. Let $\gamma = (\gamma_1, \dots, \gamma_k)'$.

Thus

$$\begin{aligned} p(D^{-1} | b, \beta, y, \tau) &= p(D^{-1} | \gamma, \beta, y, \tau) \\ &\propto |D^{-1}|^{\frac{\nu_0 + k - v - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left[(C_0^{-1} D^{-1}) + \sum_{l=1}^k \gamma_l' D^{-1} \gamma_l \right] \right\}, \end{aligned}$$

so that

$$[D^{-1} | b, \beta, y, \tau] = W_v \left(\nu_0 + k, \left(C_0^{-1} + \sum_{l=1}^k \gamma_l \gamma_l' \right)^{-1} \right). \quad (6.3)$$

The Gibbs sampler described above must now be modified by an insertion of a step to sample D^{-1} from its posterior.

One additional piece of the model is recommended by Bush and MacEachern (1996, *Biometrika*) as an aid to convergence of the Gibbs sampler. To speed mixing over the entire parameter space, they suggest moving around the γ_l 's after determining how the b_i 's are grouped.

The conditional density of γ_l is

$$p(\gamma_l \mid \beta, \tau, b, D, y) \propto \phi(\gamma_l \mid 0, D) \prod_{i \in l} p(y_i \mid \beta, b, D, \tau),$$

which implies that

$$[\gamma_l \mid \beta, \tau, b, y] \sim N\left(\tau Q_l \sum_{i \in l} Z'_i(y_i - X_i \beta), Q_l\right),$$

where

$$Q_l = \left(D^{-1} + \tau \sum_{i \in l} Z'_i Z_i\right)^{-1}.$$

The Gibbs sampling scheme is modified by inserting the following steps in between 3.n and 4:

3a.1 Sample $\gamma_1^{(i+1)}$ from $[\gamma_1 \mid \beta^{(i+1)}, b^{(i+1)}, \tau^{(i+1)}, D^{-1(i)}, y]$.

⋮

3a.k Sample $\gamma_k^{(i+1)}$ from $[\gamma_k \mid \beta^{(i+1)}, b^{(i+1)}, \tau^{(i+1)}, D^{-1(i)}, y]$.

Note that selecting a new value for γ_l changes the b_i 's for subjects in cluster l . Then the **final** Gibbs sampling scheme is:

0. Select starting values $b(0), \tau^{(0)}, D^{-1(0)}$. Set $i = 0$.
1. Sample $\beta^{(i+1)}$ from $[\beta \mid b^{(i)}, \tau^{(i)}, D^{-1(i)}, y]$ according to
 $[\beta \mid b, \tau, D, y] \sim N_p(Ta, T)$.
2. Sample $\tau^{(i+1)}$ from $[\tau \mid \beta^{(i+1)}, b^{(i)}, D^{-1(i)}, y]$ according to
 $[\tau \mid \beta, b, y] \sim \text{gamma} \left(\frac{n+\alpha_0}{2}, \frac{\lambda_0 + \sum_{i=1}^n r'_i r_i}{2} \right)$.

- 3.1 Sample $b_1^{(i+1)}$ from $[b_1 \mid \beta^{(i+1)}, b_j^{(i)}, j \neq i, D^{-1(i)}, \tau^{(i+1)}, y]$.
- 3.n Sample $b_n^{(i+1)}$ from $[b_n \mid \beta^{(i+1)}, b_j^{(i+1)}, j \neq i, D^{-1(i)}, \tau^{(i+1)}, y]$.
- 3a.1 Sample $\gamma_1^{(i+1)}$ from $[\gamma_1 \mid \beta^{(i+1)}, b^{(i+1)}, D^{-1(i)}, \tau^{(i+1)}, y]$.
- ⋮
- 3a.k Sample $\gamma_k^{(i+1)}$ from $[\gamma_k \mid \beta^{(i+1)}, b^{(i+1)}, D^{-1(i)}, \tau^{(i+1)}, y]$.
4. Sample $D^{-1(i+1)}$ from $[D^{-1} \mid \beta^{(i+1)}, \gamma^{(i+1)}, \tau^{(i+1)}, y]$ according to equation (6.3), where $\gamma^{(i+1)}$ is the γ vector sampled in steps 3a.1–3a.k.
5. Set $i = i + 1$ and return to step 1.

DP priors for the Generalized Linear Mixed Model

The MDP model can be constructed as follows. We have

$$\begin{aligned} p(y \mid \beta, b) &\propto \prod_{i=1}^n \prod_{j=1}^{n_i} p(y_{ij} \mid \beta, b_i) \\ &= \prod_{i=1}^n \prod_{j=1}^{n_i} \exp\{y_{ij}\theta_{ij} - a(\theta_{ij})\}, \end{aligned}$$

where

$$\theta_{ij} = \theta(x'_{ij}\beta + z'_{ij}b_i).$$

We take

$$\begin{aligned} \beta &\sim N_p(\mu_0, \Sigma_0) \\ b_i &\sim G \\ G &\sim DP(c_0 \cdot N_v(0, D)) \\ D^{-1} &\sim \text{Wishart}(\nu_0, C_0). \end{aligned}$$

Following similar arguments as in the normal random effects model, we can show that

$$p(\beta \mid b, y) \propto \exp \left\{ \sum_{i=1}^n \sum_{j=1}^{n_i} \log p(y_{ij} \mid \beta, b_i) - \frac{1}{2} (\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \right\} .$$

Thus,

$$\begin{aligned} p(b_i \mid \beta, y, b_j, j \neq i) &\propto \sum_{j \neq i}^n \exp \left\{ \sum_{j=1}^{n_i} \log p(y_{ij} \mid \beta, b_j) \right\} \delta_{b_j} \\ &+ \left[c_0 \int \exp \left\{ \sum_{j=1}^{n_i} \log p(y_{ij} \mid \beta, b_i) \right\} \phi(b_i \mid 0, D) db_i \right] \\ &\times \phi(b_i \mid 0, D) \prod_{j=1}^{n_i} p(y_{ij} \mid \beta, b_i) . \end{aligned}$$

When the sampling distribution is not normal, q_0 does not have a closed form for the GLMM. To avoid numerical integration or approximation, MacEachern and Müller (1998) describe a novel algorithm. Kleinman and Ibrahim (1998) adapt this algorithm to the GLMM.

Chapter 7:

Bayesian Methods for Survival Models

Bayesian Inference for Survival Models

We will first examine fully parametric and semi-parametric Bayesian Inference for **proportional hazards models**, i.e., the Cox model (Cox, 1972, *JRSS-B*). The semi-parametric approach will be of more practical interest since it will facilitate comparisons with Cox's partial likelihood (Cox, 1975, *Biometrika*).

A proportional hazards model is defined by a hazard function of the form

$$h(y, x) = h_0(y) \exp \{x'\beta\} ,$$

where $h_0(y)$ is called the **baseline hazard** function at time y , x is a $p \times 1$ vector of covariates, and β is a $p \times 1$ vector of regression coefficients. A specification of the hazard specifies the survival model.

We note that, in general,

$$h(y) = \frac{f(y)}{S(y)} , \quad S(y) = 1 - F(y) ,$$

so that

$$\begin{aligned} f(y) &= h(y) S(y) \\ S(y) &= 1 - F(y) \\ &= \exp\{-\Lambda(y)\} , \end{aligned}$$

where $\Lambda(y) = \int_0^y h(u) du$.

Therefore

$$f(y) = h(y) \exp\{-\Lambda(y)\} .$$

If we write $h(y) = h_0(y) \exp \{x'\beta\}$ (proportional hazards), we have

$$\begin{aligned} f(y) &= h_0(y) \exp \{x'\beta\} \exp \left\{ -\exp \{x'\beta\} \int_0^y h_0(u) du \right\} \\ &= h_0(y) \exp \{x'\beta - \Lambda_0(y) \exp \{x'\beta\}\} \\ &= h_0(y) \exp \{\eta - \Lambda_0(y) \exp \{\eta\}\}, \end{aligned}$$

where $\eta = x'\beta$ and $\Lambda_0(y) = \int_0^y h_0(u) du$.

Thus the survival density for any proportional hazards model takes the form

$$f(y) = h_0(y) \exp \{\eta - \Lambda_0(y) \exp \{\eta\}\},$$

where $\eta = x'\beta$ and $\Lambda_0(y) = \int_0^y h_0(u) du$.

If $h_0(y)$ is parametrically specified, then a fully parametric form of $f(y)$ is implied.

Definition 7.1

Suppose y has density

$$f(y) = \begin{cases} \alpha\gamma y^{\alpha-1} \exp\{-\gamma y^\alpha\} & \text{for } y > 0, \alpha > 0, \gamma > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This is a **Weibull** distribution with parameters (α, γ) , denoted $\text{Weibull}(\alpha, \gamma)$.

In survival analysis, it is often more convenient to write the density in terms of the parameterization $\lambda = \log(\gamma)$, leading to

$$f(y) = \alpha y^{\alpha-1} \exp\{\lambda - \exp\{\lambda\}y^\alpha\}.$$

Example 7.1: Exponential Distribution

Let

$$\begin{aligned} h_0(y) &= \lambda, \\ \Rightarrow \Lambda_0(y) &= \int_0^y \lambda \, du = \lambda y, \end{aligned}$$

and thus

$$f(y) = \lambda \exp\{\eta - \lambda y \exp\{\eta\}\}.$$

Example 7.2: Weibull Distribution

Here $\Lambda_0(y) = y^\alpha$, $\alpha > 0$, and thus

$$h_0(y) = \alpha y^{\alpha-1},$$

which leads to

$$f(y) = \alpha y^{\alpha-1} \exp\{\eta - y^\alpha \exp\{\eta\}\}.$$

Example 7.3: Extreme Value Distribution

For this distribution $\Lambda_0(y) = \exp\{\alpha y\}$, yielding

$$h_0(y) = \alpha \exp\{\alpha y\},$$

and thus

$$f(y) = \alpha \exp\{\alpha y\} \exp\{\eta - \exp\{\alpha y + \eta\}\}.$$

Bayesian analysis of fully parametric proportional hazards models proceeds as usual with prior distributions on all the parameters. In general, closed form results are **not** available for any of the posterior distributions.

For example, for the Weibull model,

$$\begin{aligned} f(y) &= \alpha y^{\alpha-1} \exp\{x'\beta - y^\alpha \exp\{x'\beta\}\}, \\ \beta &\sim N_p(\mu_0, \Sigma_0), \\ \alpha &\sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right), \end{aligned}$$

and we take β and α to be independent a priori.

The likelihood function based on n observations $y = (y_1, \dots, y_n)$ (no censoring) is given by

$$L(\beta, \alpha) = \prod_{i=1}^n p(y_i | \alpha, \beta) ,$$

where $p(y_i | \alpha, \beta) = \alpha y_i^{\alpha-1} \exp\{x_i' \beta - y_i^\alpha \exp\{x_i' \beta\}\}$ and the joint posterior of (α, β) is proportional to

$$p(\alpha, \beta | y) \propto \left\{ \prod_{i=1}^n p(y_i | \alpha, \beta) \right\} \pi(\beta) \pi(\alpha) ,$$

where

$$\pi(\beta) \propto \exp\left\{-\frac{1}{2}(\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0)\right\} ,$$

$$\pi(\alpha) \propto \alpha^{\frac{\delta_0}{2}-1} \exp\left\{-\frac{\alpha \gamma_0}{2}\right\} .$$

Typically survival data is **right censored**. We will assume a non-informative censoring mechanism.

In this case, the likelihood function for a proportional hazards model with right censored data is

$$L(\beta, h_0(y)) = \prod_{i=1}^n [h_0(y_i) \exp\{\eta_i\}]^{\nu_i} \left(S_0(y_i)^{\exp\{\eta_i\}} \right)$$

where

$$S_0(y_i) = \exp \left\{ - \int_0^{y_i} h_0(u) \, du \right\},$$

$\eta_i = x'_i \beta$, and ν_i is a censoring indicator which takes the value 1 if y_i is a failure time, and 0 if it is a censoring time.

That is,

$$\nu_i = \begin{cases} 1 & \text{not censored} \\ 0 & i^{\text{th}} \text{ observation censored} \end{cases}.$$

For the Weibull model, the likelihood becomes

$$L(\beta, \alpha) = \prod_{i=1}^n [\alpha y_i^{\alpha-1} \exp\{x_i' \beta\}]^{\nu_i} (\exp\{-y_i^\alpha \exp\{x_i' \beta\}\}) .$$

The posterior of (β, α) is given by

$$p(\beta, \alpha | y, \nu) \propto L(\beta, \alpha) \pi(\beta, \alpha),$$

where we can take

$$\pi(\beta, \alpha) = \pi(\beta) \pi(\alpha) .$$

↑ ↑
normal gamma

Exponential Model

The exponential model is the most fundamental parametric model in survival analysis.

Consider identically distributed (i.i.d.) survival times $y = (y_1, y_2, \dots, y_n)'$, each having an $\text{Exponential}(\lambda)$ distribution.

Denote the censoring indicators by $\nu = (\nu_1, \nu_2, \dots, \nu_n)'$, where $\nu_i = 0$ if y_i is right censored and $\nu_i = 1$ if y_i is a failure time.

Let $f(y_i|\lambda) = \lambda \exp\{-\lambda y_i\}$ denote the density for y_i , $S(y_i|\lambda) = \exp\{-\lambda y_i\}$ denote the survival function, and $D = (n, y, \nu)$ denote the observed data.

We can write the likelihood function of λ as

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n f(y_i|\lambda)^{\nu_i} S(y_i|\lambda)^{1-\nu_i} \\ &= \lambda^d \exp \left\{ -\lambda \sum_{i=1}^n y_i \right\}, \end{aligned}$$

where $d = \sum_{i=1}^n \nu_i$.

The conjugate prior for λ is the gamma prior. Let $\text{gamma}(\alpha_0, \lambda_0)$ denote the gamma distribution with parameters (α_0, λ_0) , with density given by

$$\pi(\lambda) \propto \lambda^{\alpha_0-1} \exp\{-\lambda_0 \lambda\}.$$

Then, taking a $\text{gamma}(\alpha_0, \lambda_0)$ prior for λ , the posterior distribution of λ is given by

$$\begin{aligned}
 p(\lambda|D) &\propto L(\lambda|D)\pi(\lambda|\alpha_0, \lambda_0) \\
 &\propto \left(\lambda^{\sum_{i=1}^n \nu_i} \exp\left\{-\lambda \sum_{i=1}^n y_i\right\} \right) \\
 &\quad \times (\lambda^{\alpha_0-1} \exp\{-\lambda_0\lambda\}) \\
 &= \lambda^{\alpha_0+d-1} \exp\left\{-\lambda(\lambda_0 + \sum_{i=1}^n y_i)\right\}. \tag{7.1}
 \end{aligned}$$

Thus we recognize the kernel of the posterior distribution in (7.1) as a $\text{gamma}(\alpha_0 + d, \lambda_0 + \sum_{i=1}^n y_i)$ distribution.

The posterior mean and variance of λ are thus given by

$$E(\lambda|D) = \frac{\alpha_0 + d}{\lambda_0 + \sum_{i=1}^n y_i}$$

and

$$\text{Var}(\lambda|D) = \frac{\alpha_0 + d}{(\lambda_0 + \sum_{i=1}^n y_i)^2}.$$

The posterior predictive distribution of a future failure time y_f is given by

$$\begin{aligned} p(y_f|D) &= \int_0^\infty p(y_f|\lambda)p(\lambda|D) d\lambda \\ &= \begin{cases} \frac{(d+\alpha_0)(\lambda_0+\sum_{i=1}^n y_i)^{(\alpha_0+d)}}{(\lambda_0+\sum_{i=1}^n y_i+y_f)^{(\alpha_0+d+1)}} & \text{if } y_f > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

To build a regression model, we introduce covariates through λ and write $\lambda_i = \varphi(x'_i \beta)$, where x_i is a $p \times 1$ vector of covariates, β is a $p \times 1$ vector of regression coefficients, and $\varphi(\cdot)$ is a known function.

A common form of φ is to take $\varphi(x'_i \beta) = \exp\{x'_i \beta\}$. Another form of φ is $\varphi(x'_i \beta) = (x'_i \beta)^{-1}$.

Using $\varphi(x'_i \beta) = \exp\{x'_i \beta\}$, we are led to the likelihood function

$$\begin{aligned}
 L(\beta) &= \prod_{i=1}^n f(y_i | \lambda_i)^{\nu_i} S(y_i | \lambda_i)^{1-\nu_i} \\
 &= \prod_{i=1}^n [\exp\{x'_i \beta\} \exp\{-y_i \exp\{x'_i \beta\}\}]^{\nu_i} \\
 &\quad \times [\exp\{-y_i \exp\{x'_i \beta\}\}]^{(1-\nu_i)} \\
 &= \exp \left\{ \sum_{i=1}^n \nu_i x'_i \beta \right\} \exp \left\{ - \sum_{i=1}^n y_i \exp\{x'_i \beta\} \right\}. \tag{7.2}
 \end{aligned}$$

In (7.2), we define $D = (n, y, X, \nu)$, where X is the $n \times p$ matrix of covariates with i^{th} row x'_i .

Common prior distributions for β include the uniform improper prior, i.e., $\pi(\beta) \propto 1$, and a normal prior.

In the regression setting, closed forms for the posterior distribution of β are generally not available, and therefore one needs to use MCMC methods.

Due to the availability of statistical software such as SAS and JAGS, the regression model in (7.2) can easily be fitted using MCMC techniques.

Suppose we specify a p -dimensional normal prior for β , denoted by $N_p(\mu_0, \Sigma_0)$, where μ_0 denotes the prior mean and Σ_0 denotes the prior covariance matrix.

Then the posterior distribution of β is given by

$$p(\beta|D) \propto L(\beta)\pi(\beta),$$

where $\pi(\beta)$ is the multivariate normal density with mean μ_0 and covariance matrix Σ_0 .

To carry out Gibbs sampling for this model, we need to write out the full conditional distributions.

Let β_j denote the j^{th} component of β , and let $\beta^{(-j)}$ denote the β vector without the j^{th} component.

Then, the j^{th} full conditional can be written as

$$p(\beta_j | D, \beta^{(-j)}) \propto L(\beta_j, \beta^{(-j)})\pi(\beta_j, \beta^{(-j)}) \quad (7.3)$$

for $j = 1, 2, \dots, p$.

To sample from each of the full conditionals in (7.3), we can use a rejection algorithm. An efficient algorithm is available here since each of the full conditionals is log-concave, and thus we can use the adaptive rejection algorithm of Gilks and Wild (1992) to sample from each of the full conditionals.

A posterior summary often of interest is the posterior distribution of the survival function at a particular value y .

For the exponential model, the survival function of an individual with covariate vector x at the point y is given by $S(y) = \exp\{-y \exp\{x'\beta\}\}$, so that if one obtains samples from the posterior distribution of β , we can readily calculate posterior summaries of $S(y)$, such as the posterior mean, variance, and credible intervals.

Example 7.4: Melanoma Data (Exponential Model)

We consider the E1684 melanoma clinical trial with $n = 285$ subjects and two treatments: high-dose IFN and observation.

We use the likelihood in (7.2) to model time until relapse with treatment as a single covariate.

Thus $\beta = (\beta_0, \beta_1)'$, where β_0 denotes the intercept term and β_1 denotes the coefficient for treatment.

A priori, we take

$$\beta \sim N_2(0, 10^4 I_2).$$

We fit the model in JAGS and in SAS using PROC MCMC.

Data in rectangular format E1684 data

faultime	rfscens	trt
1.15068	1	1
0.62466	1	1
1.89863	0	0
0.45479	1	0
2.09041	1	0
9.38356	0	1
...
...

The data set contains the following columns:

- ▶ **faultime**: failure/censoring time for each patient,
- ▶ **rfscens**: failure indicator (1 if failed, 0 if censored),
- ▶ **trt**: treatment (1 if high-dose IFN, 0 if observation).

To run the model in JAGS, we must create the following new variables:

- ▶ y : failure time for each patient (NA if censored),
- ▶ $y.cen$: censoring time if the patient was censored or the maximum observed time if the patient was uncensored,
- ▶ $is.censored$: censoring indicator (1 if censored, 0 otherwise).

We give `is.censored` a special distribution

```
is.censored[i] ~ dinterval(y[i], y.cen[i]).
```

We must initialize the missing values of y . The unobserved event times must be greater than the censoring times, so we initialize the missing values of y to values greater than the corresponding values of $y.cen$.

We write out the model in the file *e1684JAGSexpo.txt*, as shown below:

```
model{
  # Likelihood part of Bayesian inference
  for(i in 1:N){
    is.censored[i] ~ dinterval(y[i], y.cen[i])
    y[i] ~ dexp(mu[i])
    eta[i] <- beta[1] + beta[2]*trt[i]
    mu[i] <- exp(eta[i])
  }

  # Prior distributions
  for(j in 1:p){
    # Specify precision (not variance or standard deviation) in normal distribution
    beta[j] ~ dnorm(0, 0.0001)
  }
}
```

R code:

```
# Prepare input data for JAGS model - create y, y.cen, is.censored variables
e1684.90 <- read.table( "mina-e1684-e1690.txt", header = TRUE )
e1684 <- e1684.90[ e1684.90$study == 1684, c(10,11,4) ]
y.fail <- ifelse( e1684$rfscens == 1, e1684$failtime, NA )
y.cen <- ifelse( e1684$rfscens == 0, e1684$failtime, max(e1684$failtime) )
is.censored <- 1 - e1684$rfscens
data.jags <- list( y = y.fail, y.cen = y.cen, is.censored = is.censored,
                    trt = e1684$trt, N = nrow(e1684), p = 2 )

# Initialize missing values of y
y.inits <- ifelse( is.na(y.fail), y.cen, y.fail ) + 5
is.na(y.inits) <- e1684$rfscens == 1
jags.inits <- list( list(y = y.inits) )

# Set up model using JAGS ("rjags" package)
library(rjags)
jags.mod <- jags.model( file = "e1684JAGSexpo.txt", data = data.jags, inits = jags.inits,
                        n.adapt = 1000, n.chains = 1 )

# Specify names of parameters to sample
jags.vars <- "beta"
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )

# Summary of parameter estimates
summary( jags.samps )
```

Analysis in JAGS

A simple JAGS run took 2.8 seconds for 10,000 iterations after a 1000-iteration burn-in.

Posterior Summaries of β Using the Exponential Model for E1684 Data

Parameter	Mean	SD	2.5%	97.5%	Median
β_0 (int)	-1.156	0.097	-1.354	-0.973	-1.154
β_1 (trt)	-0.447	0.143	-0.726	-0.166	-0.447

The 95% credible interval for β_1 is $(-0.726, -0.166)$, indicating a treatment effect in favor of IFN.

Figure 7.1 shows trace plots of the Gibbs samples and marginal posterior densities of β using the exponential model.

The trace plots in Figure 7.1 also indicate that the Gibbs sampler is mixing well.

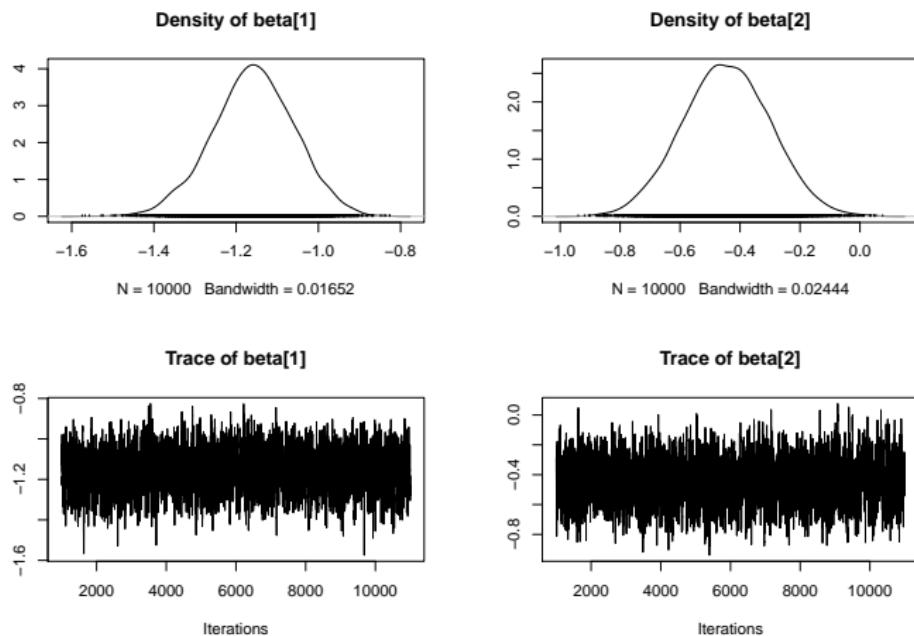


Figure 7.1: Marginal posterior densities and trace plots of β using the exponential model for E1684 data.

We now fit the model using PROC MCMC in SAS.

SAS code:

```
data e1684;
  infile "e1684SAS.dat" firstobs=2;
  input failtime rfscens trt;
run;

title 'Exponential Survival Model';
ods graphics on;
proc mcmc data=e1684 outpost=expsurvout nbi=1000 nmc=10000 seed=779;
  parms (beta0 beta1) 0;
  prior beta: ~ normal(0, var = 10000);
  /* the simplified likelihood formula is used */
  l_h = beta0 + beta1*trt;
  llike = rfscens*(l_h) - failtime*exp(l_h);
  model general(llike);
run;
ods graphics off;
```

Analysis in SAS

Posterior Summaries of β Using the Exponential Model for E1684 Data

Parameter	Mean	SD	2.5%	97.5%
β_0 (int)	-1.153	0.098	-1.350	-0.969
β_1 (trt)	-0.448	0.145	-0.723	-0.160

The 95% credible interval for β_1 is $(-0.723, -0.160)$, indicating a treatment effect in favor of IFN.

We see that the results from JAGS and SAS are consistent.

Weibull Model

The Weibull model is perhaps the most widely used parametric survival model.

Suppose we have independent identically distributed survival times $y = (y_1, y_2, \dots, y_n)'$, each having a Weibull distribution, denoted by $\text{Weibull}(\alpha, \gamma)$.

It is often more convenient to write the model in terms of the parameterization $\lambda = \log(\gamma)$, leading to

$$f(y|\alpha, \lambda) = \alpha y^{\alpha-1} \exp\{\lambda - \exp\{\lambda\}y^\alpha\}.$$

Let $S(y|\alpha, \lambda) = \exp\{-\exp\{\lambda\}y^\alpha\}$ denote the survival function.

We can write the likelihood function of (α, λ) as

$$\begin{aligned} L(\alpha, \lambda) &= \prod_{i=1}^n f(y_i|\alpha, \lambda)^{\nu_i} S(y_i|\alpha, \lambda)^{1-\nu_i} \\ &= \alpha^d \exp \left\{ d\lambda + \sum_{i=1}^n (\nu_i(\alpha - 1) \log(y_i) - \exp\{\lambda\} y_i^\alpha) \right\}, \end{aligned} \quad (7.4)$$

where $d = \sum_{i=1}^n \nu_i$.

When α is assumed known, the conjugate prior for $\exp\{\lambda\}$ is the gamma prior.

No joint conjugate prior is available when (α, λ) are both assumed unknown.

In this case, a typical joint prior specification is to take α and λ to be independent, where α has a gamma distribution and λ has a normal distribution.

Let $\text{gamma}(\alpha_0, \kappa_0)$ denote the gamma prior for α , and $N(\mu_0, \sigma_0^2)$ denote the normal prior for λ .

The joint posterior distribution of (α, λ) is given by

$$\begin{aligned} p(\alpha, \lambda | D) &\propto L(\alpha, \lambda) \pi(\alpha) \pi(\lambda) \\ &\propto \prod_{i=1}^n f(y_i | \alpha, \lambda)^{\nu_i} S(y_i | \alpha, \lambda)^{1-\nu_i} \\ &= \alpha^{\alpha_0 + d - 1} \exp \left\{ d\lambda + \sum_{i=1}^n (\nu_i(\alpha - 1) \log(y_i) - \exp\{\lambda\} y_i^\alpha) \right. \\ &\quad \left. - \kappa_0 \alpha - \frac{1}{2\sigma_0^2} (\lambda - \mu_0)^2 \right\}. \end{aligned}$$

The joint posterior distribution of (α, λ) does not have a closed form, but it can be shown that the conditional posterior distributions $[\alpha|\lambda, D]$ and $[\lambda|\alpha, D]$ are log-concave and thus Gibbs sampling is straightforward for this model.

To build the Weibull regression model, we introduce covariates through λ and write $\lambda_i = x'_i \beta$.

Common prior distributions for β include the uniform improper prior and a normal prior.

Assuming a $N_p(\mu_0, \Sigma_0)$ prior for β and a gamma prior for α , we are led to the joint posterior

$$\begin{aligned} p(\beta, \alpha | D) \propto & \alpha^{\alpha_0 + d - 1} \exp \left\{ \sum_{i=1}^n \left(\nu_i x_i' \beta + \nu_i (\alpha - 1) \log(y_i) \right. \right. \\ & \left. \left. - y_i^\alpha \exp\{x_i' \beta\} \right) - \kappa_0 \alpha \right. \\ & \left. - \frac{1}{2} (\beta - \mu_0) \Sigma_0^{-1} (\beta - \mu_0) \right\}. \end{aligned}$$

This model can be easily fit in JAGS.

Example 7.5: Mice Data

Dellaportas and Smith (1993) analyze data from Grieve (1987) on photocarcinogenicity in four groups of mice, each group containing 20 mice, who have recorded a survival time and whether they died or were censored at that time. A * indicates censoring. A portion of the data, giving survival times in weeks, are shown below.

Mouse	Irradiated control	Vehicle control	Test substance	Positive control
1	12	32	22	27
...				
18	*40	30	24	12
19	31	37	37	17
20	36	27	29	26

The survival distribution is assumed to be Weibull. That is

$$p(y_i|\beta) = \alpha \exp\{x_i'\beta\} y_i^{\alpha-1} \exp\{-\exp\{x_i'\beta\}y_i^\alpha\},$$

where y_i is the failure time of an individual with covariate vector x_i , and $\beta = (\beta_1, \dots, \beta_4)'$ is a 4×1 vector of unknown regression coefficients. This leads to a baseline hazard function of the form

$$h_0(y_i) = \alpha y_i^{\alpha-1}.$$

Setting $\mu_i = \exp\{x'_i\beta\}$ gives the parameterization

$$y_i|\beta \sim \text{Weibull}(\alpha, \mu_i) .$$

For censored observations, the survival distribution is a **truncated Weibull**, with lower bound corresponding to the censoring time.

A priori, we take

$$\beta \sim N_4(0, 10^4 I_4) ,$$

and

$$\alpha \sim \text{gamma}(1, 0.001) = \text{Exponential}(0.001) .$$

Note that the prior precision for β is $10^{-4}I_4$. The prior for α decreases slowly on the positive real line.

Median survival for individuals with covariate vector x_j is given by

$$m_j = ((\log 2) \exp\{-x'_j\beta\})^{1/\alpha} .$$

Data in rectangular format <i>mice.dat</i>			
y	cen.val	is.censored	group
12	0	0	1
17	0	0	1
21	0	0	1
25	0	0	1
...
...
35	0	0	1
NA	40	1	1
31	0	0	1
36	0	0	1
32	0	0	2
27	0	0	2
23	0	0	2
12	0	0	2
18	0	0	2
NA	40	1	2
...
...

The file *mice.dat* contains the following columns:

- ▶ **y**: failure time for each mouse (NA if censored),
- ▶ **cen.val**: censoring times for censored mice and 0 for mice with observed failure times,
- ▶ **is.censored**: censoring indicator (1 if censored, 0 otherwise),
- ▶ **group**: treatment group.

To run the model in JAGS, we must create a new variable `y.cen` which contains the censoring time if the mouse was censored or the maximum observed time (usually the end of the observation period) for uncensored mice.

We must also initialize the missing values of `y`. To do so, we initialize the missing values of `y` to values greater than the corresponding values of `y.cen`.

The contrasts, β_j with group 1 (the irradiated control), are calculated at the end. Alternatively, we could have included a grand mean term in the relative risk model and constrained β_1 to be zero.

We write out the model in the file *miceJAGS.txt*, as shown below:

```
model{  
  # Likelihood part of Bayesian inference  
  for(i in 1:N){  
    is.censored[i] ~ dinterval(y[i], y.cen[i])  
    y[i] ~ dweib(alpha, mu[i])  
    mu[i] <- exp(beta[group[i]])  
  }  
  
  # Prior distributions  
  for(j in 1:p){  
    # Specify precision (not variance or standard deviation) in normal distribution  
    beta[j] ~ dnorm(0, 0.0001)  
    median[j] <- ( log(2) * exp(-beta[j]) )^(1/alpha)  
  }  
  # Use rate parameterization for gamma distribution  
  alpha ~ dexp(0.001)  
  
  # Contrasts with group 1 (irradiated control)  
  veh.control <- beta[2] - beta[1]  
  test.sub <- beta[3] - beta[1]  
  pos.control <- beta[4] - beta[1]  
}
```

R code:

```
# Prepare input data for JAGS model - create y.cen variable
mice <- read.table( "mice.dat", header = TRUE )
y.cen <- ifelse( is.na(mice$y), mice$cen.val, max(mice$y, mice$cen.val, na.rm = TRUE) )
data.jags <- list( y = mice$y, y.cen = y.cen, is.censored = mice$is.censored,
                    group = mice$group, N = nrow(mice), p = 4 )

# Initialize missing values of y
y.inits <- ifelse( is.na(mice$y), mice$cen.val, mice$y ) + 5
is.na(y.inits) <- (1 - mice$is.censored) == 1
jags.inits <- list( list(y = y.inits) )

# Set up model using JAGS ("rjags" package)
library(rjags)
jags.mod <- jags.model( file = "miceJAGS.txt", data = data.jags, inits = jags.inits,
                        n.adapt = 1000, n.chains = 1 )

# Specify names of parameter to sample
jags.vars <- numeric(5)
jags.vars[1] <- "alpha"
jags.vars[2] <- "median"
jags.vars[3] <- "veh.control"
jags.vars[4] <- "test.sub"
jags.vars[5] <- "pos.control"
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )

# Summary of parameter estimates
summary( jags.samps )
```

Analysis in JAGS

A simple JAGS run took 2.3 seconds for 10,000 iterations after a 1000-iteration burn-in. The output is as follows:

Posterior Summaries Using the Weibull Model for Mice Data

Parameter	Mean	SD	2.5%	97.5%	Median
veh.control	-1.20	0.37	-1.95	-0.49	-1.20
test.sub	-0.36	0.35	-1.03	0.31	-0.36
pos.control	0.40	0.35	-0.29	1.08	0.40
α	3.33	0.36	2.70	4.07	3.31
median[1] (irr)	24.51	1.86	21.06	28.40	24.45
median[2] (veh)	35.27	3.05	29.94	41.91	35.07
median[3] (test)	27.30	2.19	23.36	31.92	27.19
median[4] (pos)	21.74	1.71	18.69	25.38	21.64

Example 7.6: Melanoma Data (Weibull Model)

We now consider the melanoma data using a Weibull model, whose likelihood is given by (7.4).

We use the same priors for β as before, and for α we take

$$\alpha \sim \text{gamma}(1, 0.001) = \text{Exponential}(0.001) .$$

We once again fit the model in both JAGS and SAS, and we prepare the data for JAGS in the same manner as before.

We write out the model in the file *e1684JAGSweib.txt*, as shown below:

```
model{  
  # Likelihood part of Bayesian inference  
  for(i in 1:N){  
    is.censored[i] ~ dinterval(y[i], y.cen[i])  
    y[i] ~ dweib(alpha, mu[i])  
    eta[i] <- beta[1] + beta[2]*trt[i]  
    mu[i] <- exp(eta[i])  
  }  
  
  # Prior distributions  
  for(j in 1:p){  
    # Specify precision (not variance or standard deviation) in normal distribution  
    beta[j] ~ dnorm(0, 0.0001)  
  }  
  alpha ~ dexp(0.001)  
}
```

R code:

```
# Prepare input data for JAGS model - create y, y.cen, is.censored variables
e1684.90 <- read.table( "mina-e1684-e1690.txt", header = TRUE )
e1684 <- e1684.90[ e1684.90$study == 1684, c(10,11,4) ]
y.fail <- ifelse( e1684$rfscens == 1, e1684$failtime, NA )
y.cen <- ifelse( e1684$rfscens == 0, e1684$failtime, max(e1684$failtime) )
is.censored <- 1 - e1684$rfscens
data.jags <- list( y = y.fail, y.cen = y.cen, is.censored = is.censored,
                    trt = e1684$trt, N = nrow(e1684), p = 2 )

# Initialize missing values of y
y.inits <- ifelse( is.na(y.fail), y.cen, y.fail ) + 5
is.na(y.inits) <- e1684$rfscens == 1
jags.inits <- list( list(y = y.inits) )

# Set up model using JAGS ("rjags" package)
library(rjags)
jags.mod <- jags.model( file = "e1684JAGSweib.txt", data = data.jags, inits = jags.inits,
                        n.adapt = 1000, n.chains = 1 )

# Specify names of parameters to sample
jags.vars <- numeric(2)
jags.vars[1] <- "beta"
jags.vars[2] <- "alpha"
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )

# Summary of parameter estimates
summary( jags.samps )
```

Analysis in JAGS

A simple JAGS run took 6.5 seconds for 10,000 iterations after a 1000-iteration burn-in.

Posterior Summaries of β Using the Weibull Model for E1684 Data

Parameter	Mean	SD	2.5%	97.5%	Median
α	0.586	0.035	0.518	0.656	0.586
β_0 (int)	-0.622	0.107	-0.843	-0.423	-0.617
β_1 (trt)	-0.373	0.147	-0.661	-0.083	-0.373

We see that the posterior estimates of β_0 and β_1 are similar to the exponential model, and the 95% credible interval for β_1 is $(-0.661, -0.083)$.

Figure 7.2 shows trace plots and marginal posterior densities of β_0 , β_1 , and α .

Again, the trace plots in Figure 7.2 show good mixing of the Gibbs samples for all three parameters.

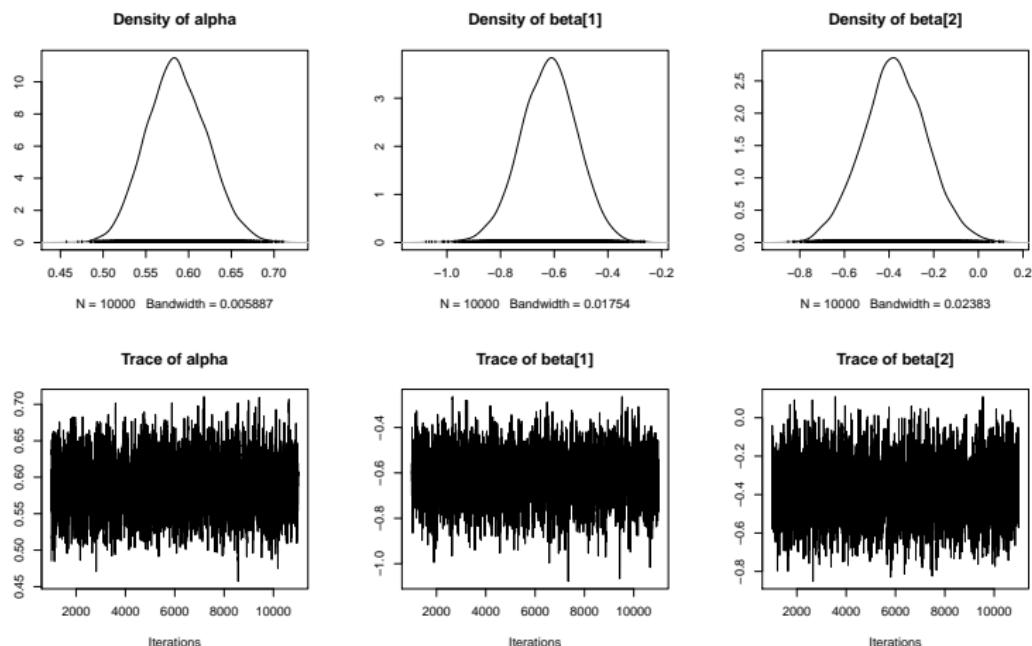


Figure 7.2: Marginal posterior densities and trace plots of (α, β) using the Weibull model for E1684 data.

We use PROC MCMC to fit the model in SAS.

SAS code:

```
data e1684;
  infile "e1684SAS.dat" firstobs=2;
  input failtime rfscens trt;
run;

title 'Weibull Survival Model';
ods graphics on;
proc mcmc data=e1684 outpost=weisurvout nbi=1000 nmc=10000 seed=779
  monitor=(_parms_);
  parms alpha 1 (beta0 beta1) 0;
  prior beta: ~ normal(0, var=10000);
  prior alpha ~ gamma(0.001,is=0.001);
  lambda = beta0 + beta1*trt;
  /* the simplified likelihood formula is used */
  llike = rfscens*(log(alpha) + (alpha-1)*log(failtime) + lambda) -
    exp(lambda)*(failtime**alpha);
  model general(llike);
run;
ods graphics off;
```

Analysis in SAS

Posterior Summaries of β Using the Weibull Model for E1684 Data

Parameter	Mean	SD	2.5%	97.5%
α	0.584	0.035	0.513	0.651
β_0 (int)	-0.615	0.108	-0.820	-0.400
β_1 (trt)	-0.378	0.150	-0.685	-0.091

We see that the posterior estimates of β_0 and β_1 are similar to the exponential model, and the 95% credible interval for β_1 is $(-0.685, -0.091)$.

As with the exponential model, we see that the results from JAGS and SAS are consistent.

Piecewise Constant Hazard Model

One of the most convenient and popular models for semiparametric survival analysis is the piecewise constant hazard model.

Semiparametric Bayesian survival analysis first involves a discretization of the time axis.

Then over each time interval, a parametric model is specified.

To construct this model, we first construct a finite partition of the time axis, $0 < s_1 < s_2 < \dots < s_J$, with $s_J > y_i$ for all $i = 1, 2, \dots, n$.

Thus, we have the J intervals $(0, s_1]$, $(s_1, s_2]$, \dots , $(s_{J-1}, s_J]$. In the j^{th} interval, we assume a constant baseline hazard $h_0(y) = \lambda_j$ for $y \in I_j = (s_{j-1}, s_j]$.

Let $D = (n, y, X, \nu)$ denote the observed data, where $y = (y_1, y_2, \dots, y_n)'$, $\nu = (\nu_1, \nu_2, \dots, \nu_n)'$ with $\nu_i = 1$ if the i^{th} subject failed and 0 otherwise, and X is the $n \times p$ matrix of covariates with i^{th} row x'_i .

Let $x'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ denote the $p \times 1$ vector of covariates for the i^{th} subject, and $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is the corresponding vector of regression coefficients.

Letting $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_J)'$, we can write the likelihood function of (β, λ) for the n subjects as

$$\begin{aligned} L(\beta, \lambda) &= \prod_{i=1}^n \prod_{j=1}^J (\lambda_j \exp\{x_i' \beta\})^{\delta_{ij} \nu_i} \exp \left\{ -\delta_{ij} \left[\lambda_j (y_i - s_{j-1}) \right. \right. \\ &\quad \left. \left. + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \right] \exp\{x_i' \beta\} \right\}, \end{aligned} \quad (7.5)$$

where $\delta_{ij} = 1$ if the i^{th} subject failed or was censored in the j^{th} interval, and 0 otherwise.

The indicator δ_{ij} is needed to properly define the likelihood over the J intervals. The semiparametric model in (7.5), sometimes referred to as a *piecewise exponential model*, is quite general and can accommodate various shapes of the baseline hazard over the intervals.

If $J = 1$, the model reduces to a parametric exponential model with failure rate parameter $\lambda \equiv \lambda_1$.

The piecewise exponential model is a useful and simple model for modeling survival data. It serves as the benchmark for comparisons with other semiparametric or fully parametric models for survival data.

A common prior of the baseline hazard λ is the independent gamma prior $\lambda_j \sim \text{gamma}(\alpha_{0j}, \lambda_{0j})$ for $j = 1, 2, \dots, J$.

Here α_{0j} and λ_{0j} are prior parameters which can be elicited through the prior mean and variance of λ_j .

Another approach is to build prior correlation among the λ_j 's using a correlated prior $\psi \sim N(\psi_0, \Sigma_\psi)$, where $\psi_j = \log(\lambda_j)$ for $j = 1, 2, \dots, J$.

Another way to correlate the λ_j 's is to specify the prior

$$\lambda_k | \lambda_{k-1} \sim \text{gamma}(\alpha_k, \alpha_k / \lambda_{k-1})$$

for $k > 1$, $k = 2, \dots, J$, and take $\lambda_1 \sim \text{gamma}(\alpha_1, \gamma_1)$.

We can specify a normal (or uniform) prior for β and take β and λ independent a priori.

Example 7.7: Melanoma Data (Piecewise Exponential Model)

We now consider the melanoma data using a piecewise exponential model, whose likelihood is given by (7.5).

We choose $J = 5$, and we determine the interval endpoints such that each interval includes approximately the same number of failures.

For the treatment coefficient β , we take the prior to be

$$\beta \sim N(0, 10^4).$$

For the priors on the baseline hazards, we take

$$\lambda_j \sim \text{gamma}(0.01, 0.01), \quad j = 1, \dots, J.$$

We fit the model using JAGS, Stan, and SAS.

For the model in SAS, we organize the data to include three columns: (1) the failure/censoring time, (2) the censoring indicator, and (3) the treatment indicator (see data structure on page 747).

For JAGS and Stan, we fit the equivalent Poisson model. See Holford (1980, Biometrics) and Laird and Oliver (1981, JASA).

The equivalent Poisson model requires the data to be in long format where each patient has a separate observation for each interval prior to and including the interval in which the patient's failure/censored time was observed. For example, if a patient has an observed time in the third interval, then they would have three separate observations.

In long format, the response for the i^{th} patient in the j^{th} interval is defined as

$$y_{ij} = \begin{cases} s_j - s_{j-1} & \text{if the patient survived past interval } j \\ y_i - s_{j-1} & \text{if the patient's observed time is in interval } j \end{cases},$$

where y_i is the patient's observed time and s_j is the right endpoint of the j^{th} interval, $i = 1, \dots, n$, $j = 1, \dots, J$.

We write out the model in the file *e1684JAGSpe.txt*, as shown below:

```
model{  
  # Likelihood part of Bayesian inference  
  for(i in 1:Nobs){  
    d[i] ~ dpois(mu[i])  
    log(mu[i]) <- theta[i]  
    theta[i] <- log(yint[i]) + log(lambda[int[i]]) + beta.trt * trt[i]  
  }  
  
  # Prior distributions  
  for(j in 1:J){  
    # Use rate parameterization for gamma distribution  
    lambda[j] ~ dgamma(0.01, 0.01)  
  }  
  # Specify precision (not variance or standard deviation) in normal distribution  
  beta.trt ~ dnorm(0, 0.0001)  
}
```

R code to run JAGS code:

```

# Read in data
e1684.90 <- read.table( "mina-e1684-e1690.txt", header = TRUE )
e1684 <- e1684.90[ e1684.90$study == 1684, c(10,11,3:5) ]
e1684 <- e1684[ e1684$faultime != 0, ]      # remove observation with observed time of 0

# Pick intervals with equal number of failures in each interval, J = 5
int.right <- c( quantile( e1684$faultime[e1684$rfscens == 1], probs = c(.2, .4, .6, .8) ),
               max( e1684$faultime ) + 1 )

# Format data into long format such that each patient has a separate observation for each interval
library(tidyverse)
library(survival)
e1684.long <- survSplit( formula = Surv(faultime, rfscens) ~ ., data = e1684,
                           cut = int.right, start = "ystart", id = "id" ) %>%
  group_by(group = cumsum(id != lag(id, default = first(id)))) %>%
  mutate(interval = factor(ystart), interval_length = faultime - ystart, which.int = row_number()) %>%
  ungroup() %>% select(-group) %>% as_tibble
data.jags <- list( yint = e1684.long$interval_length, int = e1684.long$which.int,
                    d = e1684.long$rfscens, trt = e1684.long$trt,
                    Nobs = nrow(e1684.long), J = length(int.right) )

# Set up model using JAGS ("rjags" package)
library(rjags)
jags.mod <- jags.model( file = "e1684JAGSspe.txt", data = data.jags,
                        n.adapt = 1000, n.chains = 1 )

# Specify names of parameters to sample
jags.vars <- c( "lambda", "beta.trt" )
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )

# Summary of parameter estimates
summary( jags.samps )

```

For the Stan model, we specify uniform priors for β and $\log(\lambda_j)$.

R code to run Stan code:

```
# Read in data
e1684.90 <- read.table( "mina-e1684-e1690.txt", header = TRUE )
e1684 <- e1684.90[ e1684.90$study == 1684, c(10,11,3:5) ]
e1684 <- e1684[ e1684$faultime != 0, ]      # remove observation with observed time of 0

# Pick intervals with equal number of failures in each interval, J = 5
int.right <- c( quantile( e1684$faultime[e1684$rfscens == 1], probs = c(.2, .4, .6, .8) ),
               max( e1684$faultime ) + 1 )

# Format data into long format such that each patient has a separate observation for each interval
library(tidyverse)
library(survival)
e1684.long <- survSplit( formula = Surv(faultime, rfscens) ~ ., data = e1684,
                           cut = int.right, start = "ystart", id = "id" ) %>%
  group_by(group = cumsum(id != lag(id, default = first(id)))) %>%
  mutate(interval = factor(ystart), interval_length = faultime - ystart, which.int = row_number()) %>%
  ungroup() %>% select(-group) %>% as_tibble

# Fit model using Stan
library(rstanarm)
stan.mod <- stan_glm( formula = rfscens ~ -1 + interval + trt + offset(log(interval_length)),
                      refresh = 0, data = e1684.long, family = poisson(link = "log"), seed = 779,
                      prior = NULL, chains = 1, iter = 10000, warmup = 1000,
                      control = list(max_treedepth = 15) )

# Posterior means, standard errors
post.samples <- as.matrix(stan.mod)
post.samples2 <- cbind( post.samples[,6], exp( post.samples[,1:5] ) )
colnames(post.samples2) <- c( "trt", "lambda1", "lambda2", "lambda3", "lambda4", "lambda5" )
colMeans(post.samples2)
apply( post.samples2, 2, sd )
```

SAS code:

```
/* Read in data */
data e1684;
  infile "e1684SAS.dat" firstobs=2;
  input failtime rfscens trt;
run;

/* Piecewise exponential model - switch reference level for trt to 0 instead of SAS default */
proc phreg data=e1684;
  class trt(ref="0");
  model failtime*rfscens(0) = trt;
  bayes seed=779 piecewise=hazard(interval=(0.19452 ,0.41644, 0.86301, 1.72055)
    prior=gamma(shape=0.01 iscale=0.01) coeffprior=normal(var=10000);
    ** right side of five intervals are 0.19452 ,0.41644, 0.86301, 1.72055, Inf;
run;
```

Posterior Summaries Using the Piecewise Exponential Model for E1684 Data

Parameter	JAGS	Stan	SAS
β (trt)	-0.367 (0.145)	-0.363 (0.143)	-0.367 (0.143)
λ_1	0.929 (0.158)	0.925 (0.154)	0.883 (0.151)
λ_2	0.916 (0.162)	0.918 (0.160)	0.944 (0.163)
λ_3	0.585 (0.101)	0.585 (0.102)	0.586 (0.102)
λ_4	0.382 (0.066)	0.383 (0.067)	0.383 (0.067)
λ_5	0.100 (0.018)	0.100 (0.018)	0.103 (0.018)

From the table, we see that all three models produced similar posterior estimates.

The likelihood in (7.5) is based on continuous survival data as opposed to grouped survival data. We use the actual survival times in the construction of the likelihood.

Grouped data likelihood is based on counting the number of survival times that fall into an interval and then treating the data as binomial.

For grouped survival data

$$\begin{aligned} P(y_i \in I_j | \lambda) &= \exp \left\{ -\exp\{x_i' \beta\} \sum_{k=1}^{j-1} \Delta_k \lambda_k \right\} \\ &\quad \times [1 - \exp\{-\Delta_j \lambda_j \exp\{x_i' \beta\}\}]^{\nu_i}. \end{aligned}$$

The likelihood function based on grouped survival data is given by

$$L(\beta, \lambda) \propto \prod_{j=1}^J G_j^*,$$

$$\begin{aligned} G_j^* &= \exp \left\{ -\lambda_j \Delta_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j - \mathcal{C}_j} \exp \{x_k' \beta\} \right\} \\ &\quad \times \prod_{l \in \mathcal{D}_j} [1 - \exp \{-\lambda_j \Delta_j \exp \{x_l' \beta\}\}], \end{aligned} \tag{7.6}$$

where $\Delta_j = s_j - s_{j-1}$ and \mathcal{R}_j , \mathcal{D}_j , and \mathcal{C}_j are the sets of patients at risk, patients having failures, and patients having censored times, respectively, in the j^{th} interval.

Example 7.8: Melanoma Data Using the Power Prior

We now consider an analysis by Chen, Harrington, and Ibrahim (2009, JRSS-C) using the E1690 melanoma data.

The E1684 data are used as the historical data (D_0), and it is incorporated via the power prior.

The power prior for (β, λ) for model (7.5) is given by

$$\pi(\beta, \lambda | a_0, D_0) \propto L(\beta, \lambda | D_0)^{a_0} \pi_0(\beta, \lambda),$$

where $\pi_0(\beta, \lambda)$ is the initial prior distribution for (β, λ) , and $L(\beta, \lambda | D_0)$ is the likelihood function in (7.5) with D_0 in place of D .

For the initial prior $\pi_0(\beta, \lambda)$, we assume that β and λ are independent, where β has a uniform prior and λ has a Jeffreys's prior.

This leads to the joint initial improper prior

$$\pi_0(\beta, \lambda) \propto \prod_{j=1}^J \lambda_j^{-1}. \quad (7.7)$$

We consider the treatment covariate alone in the example here, and thus β is one dimensional.

Table 7.1: Posterior Estimates of Hazard Ratio for E1684 Data

a_0	Posterior HR	Posterior SD	95% HPD
0	1.30	0.17	(0.99, 1.64)
0.05	1.30	0.16	(0.99, 1.63)
0.30	1.33	0.15	(1.03, 1.63)
1	1.36	0.13	(1.11, 1.62)

Table 7.1 shows results based on several values of a_0 using the initial prior in (7.7).

In Table 7.1, HR denotes the posterior hazard ratio of OBS to IFN, SD denotes the posterior standard deviation, and 95% HPD denotes 95% Highest Posterior Density intervals.

We see from Table 7.1 that as more weight is given to the historical data, the posterior hazard ratios increase and the HPD intervals become narrower and do not include 1.

This is reasonable since the posterior hazard ratios based on the E1684 data alone were much larger than E1690 alone, and therefore as more weight is given to E1684, the greater the posterior hazard ratios and the narrower the HPD intervals.

Thus, the incorporation of E1684 into the current analysis via the power prior sharpens the assessment between IFN and OBS and leads to more definitive conclusions about the effect of IFN.

This example thus demonstrates the effect of incorporating historical data into an analysis.

Gamma Process on the Cumulative Baseline Hazard

Although fully parametric survival analysis may be useful, it is typically **not** done in practice.

In practice, Cox's partial likelihood has received a lot of attention, since it eliminates the baseline hazard $h_0(y)$ from the "likelihood." Cox's partial likelihood for β takes the form

$$L(\beta) = \prod_{i=1}^r \left\{ \frac{\exp\{x_i' \beta\}}{\sum_{j \in R(y_i)} \exp\{x_j' \beta\}} \right\}, \quad (7.8)$$

where r = number of events (deaths) and $R(y_i)$ is the set of individuals at risk just prior to y_i . Estimation of β proceeds from $L(\beta)$ above.

Thus from a Bayesian standpoint, to facilitate comparisons with $L(\beta)$ above, it is of interest to take a semi-parametric approach in which a fully parametric prior is specified for the regression coefficients and a non-parametric prior is specified for the hazard rate $h_0(y)$ **or** the cumulative hazard $\Lambda_0(y) = \int_0^y h_0(u) du$. We shall discuss both approaches.

The typical prior distribution that is specified for $h_0(y)$ or $\Lambda_0(y)$ is called a **gamma process prior**.

Definition 7.2: Gamma Process

Suppose $x \sim \text{gamma}(\alpha, \lambda)$, so that

$$f(x | \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\lambda x\}, \quad x > 0, \alpha > 0, \lambda > 0.$$

Let $\alpha(s)$, $s \geq 0$ be a non-decreasing left continuous function such that $\alpha(0) = 0$, and let $z(s)$, $s \geq 0$, be a stochastic process such that

1. $z(0) = 0$
2. $z(s)$ has independent increments
3. for $s > t$, $z(s) - z(t) \sim \text{gamma}(\alpha(s) - \alpha(t), 1)$.

The independent increments assumption means that

$$\{z(s_1), z(s_2) - z(s_1), z(s_3) - z(s_2) \dots\}$$

are independent. The stochastic process $\{z(s) : s \geq 0\}$ defined this way is called a **gamma process**. We shall first discuss gamma process priors for the cumulative baseline hazard, $\Lambda_0(s)$.

It will also be convenient to represent the likelihood function for a proportional hazards model using **counting process** notation. The counting process notation can be constructed as follows:

1. Let $N_i(s) =$ number of failures which have occurred up to time s , $i = 1, \dots, n$, where $n =$ number of subjects.
2. Define the **intensity process** $I_i(s)$ as

$$I_i(s) = Y_i(s) h_0(s) \exp\{x_i' \beta\},$$

where $Y_i(s)$ is an **observed** process taking the value 1 or 0 according to whether subject i is observed at a time $\geq s$.

3. $dN_i(s) =$ increment of N_i over the time interval $(s, s + ds]$
4. The data are denoted by $D = \{N_i(s), Y_i(s), x_i, i = 1, \dots, n\}$.
The unknown parameters are β and

$$\Lambda_0(s) = \int_0^s h_0(u) du.$$

Consider the intervals $(0, s_1], (s_1, s_2], (s_2, s_3], \dots (s_{J-1}, s_J]$, where $s_J > y_i$ for all $i = 1, \dots, n$.

For the time interval sizes, we consider intervals in which at least one event occurred (either failure or censored observation).

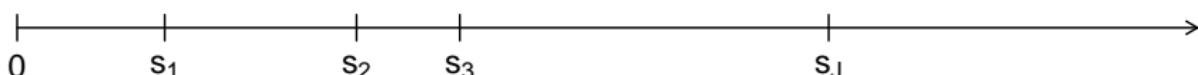


Figure 7.3:

Further let

$$dN_{ij} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ subject failed in } (s_{j-1}, s_j] \\ 0 & \text{otherwise} \end{cases},$$

$$Y_{ij} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ subject was observed in a time } \geq s_{j-1} \\ 0 & \text{otherwise} \end{cases},$$

and

$$I_{ij} = Y_{ij} \exp\{x_i' \beta\} d\Lambda_{0j},$$

where $d\Lambda_{0j}$ is the change in the cumulative baseline hazard from $(s_{j-1}, s_j]$. That is,

$$d\Lambda_{0j} = \Lambda_0(s_j) - \Lambda_0(s_{j-1}).$$

We can now write the proportional hazards model as

$$\begin{aligned}
 L(\beta, \Lambda_0(\cdot)) &= \prod_{i=1}^n \prod_{j=1}^J (I_{ij})^{dN_{ij}} \exp \left\{ -Y_{ij} \exp\{x_i' \beta\} \int_{s_{j-1}}^{s_j} h_0(u) du \right\} \\
 &= \prod_{i=1}^n \prod_{j=1}^J (I_{ij})^{dN_{ij}} \exp\{-Y_{ij} \exp\{x_i' \beta\} d\Lambda_{0j}\} \\
 &= \prod_{i=1}^n \prod_{j=1}^J (I_{ij})^{dN_{ij}} \exp\{-I_{ij}\} ,
 \end{aligned}$$

where $I_{ij} = Y_{ij} \exp\{x_i' \beta\} d\Lambda_{0j}$.

We recognize the dN_{ij} 's as independent Poisson random variables with means I_{ij} . That is,

$$dN_{ij} \sim \text{Poisson}(I_{ij}). \quad (7.9)$$

The **gamma process prior** for $\Lambda_0(s)$ implies

$$d\Lambda_{0j} \sim \text{gamma}(c_0 d\Lambda_{0j}^*, c_0).$$

That is, the increments $d\Lambda_{0j}$ are distributed as gamma random variables. Here, $\Lambda_0^*(s)$ is the prior guess for $\Lambda_0(s)$. We write

$$\Lambda_0(s) \sim GP(c_0 \Lambda_0^*(s), c_0),$$

where GP = gamma process.

The hyperparameter c_0 reflects the degree of prior belief in $\Lambda_0^*(s)$. It is convenient to take

$$d\Lambda_{0j}^* = r_0(s_j - s_{j-1}) ,$$

where r_0 is a guess at the failure rate per unit time.

Let $r_j = d\Lambda_{0j}$. It follows that

$$r_j \sim \text{gamma}(c_0 r_0(s_j - s_{j-1}), c_0), \quad j = 1, \dots, J.$$

Thus the r_j 's are prior parameters that reflect our prior beliefs in the increments of the cumulative baseline hazard, and c_0 is a scalar prior parameter that reflects our degree of prior belief in these increments.

The prior specification for β is a parametric one, and typically we choose a normal prior. That is, we take

$$\beta \sim N_p(\mu_0, \Sigma_0).$$

Let $r = (r_1, \dots, r_J)$. We write the posterior distribution as

$$p(\beta, r | D) \propto \left(\prod_{i=1}^n \prod_{j=1}^J (I_{ij})^{dN_{ij}} \exp\{-I_{ij}\} \right) \pi(\beta) \pi(r),$$

where

$$\pi(\beta) \propto \exp \left\{ -\frac{1}{2} (\beta - \mu_0)' \Sigma_0^{-1} (\beta - \mu_0) \right\}$$

and

$$\pi(r) \propto \prod_{j=1}^J r_j^{c_0 r_0 (s_j - s_{j-1}) - 1} \exp\{-r_j c_0\}.$$

Exercise 7.1

Write the likelihood function of $(\beta, \Lambda_0(\cdot))$ in

- i) counting process notation
- ii) non-counting process notation

Example 7.9: Leukemia Data (Counting Process)

We use counting process notation to analyze leukemia data from a clinical trial that compared time-to-remission in $N = 42$ patients who received one of two drug regiments: 6-MP or placebo.

We use the Poisson likelihood in (7.9) to model the counting process jumps dN_{ij} , $i = 1, \dots, N$, $j = 1, \dots, J$.

We set $J = 17$, the number of unique failure times.

Let β denote the coefficient for treatment, and a priori we take

$$\beta \sim N(0, 10^4).$$

Leukemia Data		
failtime	fail	trt
1	1	0
1	1	0
2	1	0
2	1	0
3	1	0
4	1	0
...
...
32	0	1
34	0	1
35	0	1

The data set contains the following columns:

- ▶ **failtime**: failure/censoring time for each patient,
- ▶ **fail**: failure indicator (1 if failed, 0 if censored),
- ▶ **trt**: treatment (1 if 6-MP, 0 if placebo).

We write out the model in the file *leukJAGS.txt*, as shown below:

```
model{  
  # Model  
  c0 <- .001  
  r0 <- 0.1  
  for(j in 1:J){  
    for(i in 1:N){  
      # Poisson likelihood "trick"  
      dN[i,j] ~ dpois(Idt[i,j])  
      # Intensity  
      Idt[i,j] <- Y[i,j] * exp(beta * trt[i]) * dL0[j]  
    }  
    # Gamma process prior on the cumulative baseline hazard rate  
    dL0[j] ~ dgamma(mu[j], c0)  
    mu[j] <- dL0.star[j] * c0      # prior mean cumulative hazard  
  }  
  
  for(j in 1:J){  
    dL0.star[j] <- r0 * (s[j+1] - s[j])  
  }  
  beta ~ dnorm(0, 0.0001)  
}
```

Before we run the model in JAGS, we set up the data using counting process notation.

R code:

```
# Prepare input data for JAGS model
leuk <- read.table( "leuk.dat", header = TRUE )
# s: unique failure times + maximum censoring time
s <- c( sort( unique( leuk$faultime[leuk$fail == 1] ) ), max(leuk$faultime) )
J <- length( unique( leuk$faultime[leuk$fail == 1] ) ) # number of unique failure times
N <- nrow(leuk)

# Set up data
Y <- matrix( 0, nrow = N, ncol = J )
dN <- matrix( 0, nrow = N, ncol = J )
for(i in 1:N){
  for(j in 1:J){
    # Risk set
    Y[i,j] <- ifelse( leuk$faultime[i] - s[j] >= 0, 1, 0 )
    # Counting process jump = 1 if faultime in [ s[j], s[j+1] ), i.e., if s[j] <= faultime < s[j+1]
    dN[i,j] <- Y[i,j] * ifelse( s[j+1] - leuk$faultime[i] >= 0, 1, 0 ) * leuk$fail[i]
  }
}
data.jags <- list( Y = Y, dN = dN, s = s, trt = leuk$trt, N = nrow(leuk), J = J )
```

R code (cont.):

```
# Set up model using JAGS ("rjags" package)
library(rjags)
jags.mod <- jags.model( file = "leukJAGS.txt", data = data.jags, n.adapt = 1000, n.chains = 1 )

# Specify names of parameters to sample
jags.vars <- "beta"
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )

# Summary of parameter estimates
summary( jags.samps )
```

The posterior mean for β is -1.435 with a 95% credible interval $(-2.043, -0.866)$, indicating a treatment effect in favor of 6-MP.

Models Using a Gamma Process

As an alternative to the piecewise constant model, we can specify priors on the increments in the cumulative hazard. (Kalbfleish, 1978).

We demonstrate the gamma process model with the grouped data likelihood.

Again, we construct a finite partition of the time axis, $0 < s_1 < s_2 < \dots < s_J$, with $s_J > y_i$ for all $i = 1, \dots, n$.

Thus, we have the J disjoint intervals $(0, s_1]$, $(s_1, s_2]$, \dots , $(s_{J-1}, s_J]$, and let $I_j = (s_{j-1}, s_j]$.

The observed data D is assumed to be available as grouped within these intervals, such that

$$D = (X, \mathcal{R}_j, \mathcal{D}_j, \mathcal{C}_j : j = 1, 2, \dots, J),$$

where \mathcal{R}_j , \mathcal{D}_j , and \mathcal{C}_j are the risk set, the failure set, and the censored set, respectively, of the j^{th} interval I_j .

Let h_j denote the increment in the cumulative baseline hazard in the j^{th} interval, that is

$$h_j = H_0(s_j) - H_0(s_{j-1}), \quad j = 1, 2, \dots, J.$$

The gamma process prior on the h_j 's assumes that the h_j 's are independent and

$$h_j \sim \text{gamma}(\alpha_{0j} - \alpha_{0,j-1}, c_0),$$

where $\alpha_{0j} = c_0 H^*(s_j)$, c_0 is a scalar, and H^* is a parametric guess for H_0 .

For example, if H^* is Weibull, then $H^*(s_j) = \eta_0 s_j^{\kappa_0}$, where (η_0, κ_0) are specified hyperparameters.

Thus, the hyperparameters (H^*, c_0) for h_j consist of a specified parametric cumulative hazard function $H^*(y)$ evaluated at the endpoints of the time intervals, and a positive scalar c_0 quantifying the degree of prior confidence in $H^*(y)$.

Thus, the grouped data representation can be obtained as

$$\begin{aligned} P(y_i \in I_j | h) &= \exp \left\{ -\exp\{x_i' \beta\} \sum_{k=1}^{j-1} h_k \right\} \\ &\quad \times [1 - \exp\{-h_j \exp\{x_i' \beta\}\}]^{\nu_i}, \end{aligned}$$

where $h = (h_1, h_2, \dots, h_J)'$.

This leads to the grouped data likelihood function

$$L(\beta, h|D) \propto \prod_{j=1}^J G_j,$$

$$\begin{aligned} G_j &= \exp \left\{ -h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j - \mathcal{C}_j} \exp\{x'_k \beta\} \right\} \\ &\quad \times \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp\{x'_l \beta\}\}]. \end{aligned}$$

A typical prior for β is a $N_p(\mu_0, \Sigma_0)$ distribution.

Thus, the joint posterior of (β, h) can be written as

$$\begin{aligned}\pi(\beta, h|D) &\propto \prod_{j=1}^J \left[G_j h_j^{(\alpha_{0j} - \alpha_{0,j-1})-1} \exp\{-c_0 h_j\} \right] \\ &\quad \times \exp \left\{ -\frac{1}{2} (\beta - \mu_0) \Sigma_0^{-1} (\beta - \mu_0) \right\}.\end{aligned}$$

To sample from the joint posterior distribution of (β, h) , it can be shown that $[\beta|h, D]$ is log-concave in β and thus the adaptive rejection algorithm can be used efficiently to sample the components of β .

Moreover, $[h|\beta, D]$ is also log-concave in the components of h .

We can thus carry out the following Gibbs sampling scheme:

- (i) Sample from

$$\pi(\beta_q | \beta^{(-q)}, h, D) \propto \prod_{j=1}^J G_j \exp \left\{ -\frac{1}{2} (\beta - \mu_0) \Sigma_0^{-1} (\beta - \mu_0) \right\},$$

using the adaptive rejection algorithm for $q = 1, 2, \dots, p$.

(ii) Sample from

$$\begin{aligned}\pi(h_j | h^{(-j)}, \beta, D) &\propto h_j^{(\alpha_{0j} - \alpha_{0,j-1}) - 1} \\ &\times \exp \left\{ -h_j \left(\sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp\{x'_k \beta\} + c_0 \right) \right\} \\ &\times \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp\{x'_l \beta\}\}]\end{aligned}\quad (7.10)$$

where $h^{(-j)}$ denote the h vector without the j^{th} component.

The full conditional distribution in (7.10) can be well approximated by a gamma distribution, and thus a more efficient Gibbs sampling scheme would be to replace (7.10) by

- (ii*) Sample from $[h|\beta, D]$ using independent samples from a conditional posterior approximated by

$$h_j \sim \text{gamma} \left(\alpha_{0j} - \alpha_{0,j-1} + d_j, c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp\{x'_k \beta\} \right).$$

Gamma Process on the Baseline Hazard

The assumption of independence in the increments of the cumulative baseline hazard may not be realistic or reasonable in many situations. Thus a GP prior for the cumulative baseline hazard may not be the most attractive approach. In addition, prior information on the cumulative baseline hazard may not be readily available or may be difficult to specify.

For these reasons, it may be more desirable to specify a gamma process prior on the baseline hazard rate itself rather than the cumulative baseline hazard. Thus, we can take

$$h_0(\cdot) \sim GP(\alpha(\cdot), \lambda(\cdot)) .$$

In particular, we discretize the time axis $0 < s_1 < s_2 \dots < s_J$ with $s_J > y_i$ for all $i = 1, \dots, n$, and define

$$\delta_j = h_0(s_j) - h_0(s_{j-1}) , \quad j = 1, \dots, J.$$

Thus δ_j denotes the increment in the baseline hazard in the interval $(s_{j-1}, s_j]$, $j = 1, \dots, J$.

Thus we take

$$\delta_j \sim \text{gamma}(\alpha_j, \lambda_j), \quad j = 1, \dots, J.$$

Common choices of α_j are $\alpha_j = \alpha_0 (s_j - s_{j-1})$, and common choices of λ_j are $\lambda_j = \lambda_0$.

The (Approximate) Likelihood Function

The cumulative distribution function for the proportional hazards model at time s is given by

$$\begin{aligned} F(s) &= 1 - \exp \left\{ -\exp\{\eta\} \int_0^s h_0(u) \, du \right\} \\ &\approx 1 - \exp \left\{ -\exp\{\eta\} \sum_{j=1}^J \delta_j (s - s_{j-1})^+ \right\}, \end{aligned}$$

where $(s - s_{j-1})^+ = s - s_{j-1}$ if $s - s_{j-1} > 0$, and 0, otherwise. Now let p_j denote the probability of a failure in the interval $(s_{j-1}, s_j]$, $j = 1, \dots, J$.

We have

$$\begin{aligned} p_j &= F(s_j) - F(s_{j-1}) \\ &\approx \exp \left\{ -\exp\{\eta\} \sum_{k=1}^{j-1} \delta_k (s_{j-1} - s_{k-1}) \right\} \\ &\quad \times \left[1 - \exp \left\{ -\exp\{\eta\} (s_j - s_{j-1}) \sum_{k=1}^j \delta_k \right\} \right]. \end{aligned}$$

Thus in the j^{th} interval $(s_{j-1}, s_j]$, the contribution to the likelihood function for an “exact” observation (i.e., a failure) is p_j and $1 - F(s_j)$ for a right censored observation.

Now let d_j = number of failures in the j^{th} interval, and let c_j = number of censored observations in the j^{th} interval. For ease of exposition, we order the observations so that in the j^{th} interval, the first d_j are failures and the remaining c_j are censored, $j = 1, \dots, J$. Let x_{jk} denote the vector of covariates for the k^{th} individual in the j^{th} interval, and define

$$u_{jk}(\beta) = \exp\{x'_{jk}\beta\} ,$$

$$a_j = \sum_{i=j+1}^J \sum_{k=1}^{d_i} u_{ik}(\beta)(s_{i-1} - s_{j-1}) ,$$

$$b_j = \sum_{i=j}^J \sum_{k=d_i+1}^{d_i+c_i} u_{ik}(\beta)(s_i - s_{j-1}) ,$$

$$\begin{aligned} T_j(\Delta) &= (s_j - s_{j-1}) \sum_{i=1}^j \delta_i , \\ \Delta &= (\delta_1, \dots, \delta_J). \end{aligned}$$

Then the (approximate) likelihood function of (β, Δ) is given by

$$\begin{aligned} L(\beta, \Delta) &= \left[\prod_{j=1}^J \exp\{-\delta_j(a_j + b_j)\} \right] \\ &\times \left[\prod_{j=1}^J \prod_{k=1}^{d_j} (1 - \exp\{-u_{jk}(\beta)T_j(\Delta)\}) \right]. \end{aligned}$$

Prior Distributions

Again, we assume that (β, Δ) are independent a priori, so that

$$\begin{aligned} \pi(\beta, \Delta) &= \pi(\beta)\pi(\Delta) \\ &= \pi(\beta) \prod_{j=1}^J \pi(\delta_j). \end{aligned}$$

We can take $\beta \sim N_p(\mu_0, a_0^{-1}\Sigma_0)$, and the δ_j 's are independent gamma variates.

For elicitation of (μ_0, Σ_0) , if a previous study exists with data $D_0 = (n_0, y_0, X_0, \nu_0)$, then we can construct Cox's partial likelihood based on D_0 , as

$$L^*(\beta) = \prod_{j=1}^{f_0} \left(\frac{\exp\{x'_{0j}\beta\}}{\sum_{\ell \in R(y_{0j})} \exp\{x'_{0\ell}\beta\}} \right),$$

where f_0 = number of failures in the historical data, ν_0 is the censoring indicator for the historical data, y_0 are the failure/censoring times for the historical data, X_0 is the covariate matrix, and n_0 is the sample size.

For the prior mean of β , we take μ_0 to be the solution to

$$\frac{\partial \log[L^*(\beta)]}{\partial \beta_i} = 0,$$

and we take Σ_0^{-1} to be

$$\Sigma_0^{-1} = \left[\frac{-\partial^2 \log L^*(\beta)}{\partial \beta_j \partial \beta_i} \right] \Big|_{\beta=\mu_0}.$$

Thus we use Cox's partial likelihood on the historical data to elicit (μ_0, Σ_0) . The prior parameters for Δ can come from an estimate of the hazard rate based on the historical data. We note that in model selection problems, Δ is typically viewed as a nuisance parameter and thus **vague** gamma priors are typically specified.

A straightforward Gibbs sampling scheme can be devised for sampling from the joint posterior distribution of (β, Δ) . The Gibbs sampling strategy involves the introduction of **latent variables**. The posterior density of $[\Delta|\beta, D]$, where D denotes the data for current study, is given by

$$\begin{aligned} p(\Delta|\beta, D) &\propto \left(\prod_{j=1}^J \exp\{-\delta_j(a_j + b_j)\} \right) \\ &\times \left(\prod_{j=1}^J \prod_{k=1}^{d_j} (1 - \exp\{-u_{jk}(\beta)T_j(\Delta)\}) \right) \\ &\times \left(\prod_{j=1}^J \pi(\delta_j) \right). \end{aligned}$$

We define **latent variables** in order to make the components of Δ **independent** a posteriori. We do this by first defining

$$e_j = (e_{j1}, \dots, e_{jd_j}), \quad j = 1, \dots, J,$$

to be **independent exponential random variables truncated at 1**, with mean equal to

$$(T_j(\Delta) u_{jk}(\beta))^{-1}.$$

Thus

$$p(e_{jk}) = \begin{cases} \frac{T_j(\Delta) u_{jk}(\beta)}{1 - \exp\{-T_j(\Delta) u_{jk}(\beta)\}} \exp\{-e_{jk} T_j(\Delta) u_{jk}(\beta)\}, & e_{jk} \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Letting $e = (e_1, \dots, e_J)$, we can write the posterior distribution of $[\Delta | \beta, e, D]$ as

$$\begin{aligned} p(\Delta | \beta, e, D) &\propto \left(\prod_{j=1}^J (T_j(\Delta))^{d_j} \right) \left(\exp \left\{ -\sum_{j=1}^J \sum_{k=1}^{d_j} e_{jk} T_j(\Delta) u_{jk}(\beta) \right\} \right) \\ &\times \left(\prod_{j=1}^J \exp \{-\delta_j(a_j + b_j)\} \right) \left(\prod_{j=1}^J \pi(\delta_j) \right). \end{aligned}$$

Now we consider an additional set of latent variables $q_j = (q_{j1}, \dots, q_{jj})$, $j = 1, \dots, J$, where the q_j 's are independent multinomials. Each q_j is an j -cell multinomial of d_j independent trials with probability of the k^{th} cell defined to be

$$p_k = \frac{\delta_k}{\sum_{i=1}^j \delta_i}.$$

Letting $q = (q_1, \dots, q_J)$, we are led to

$$\begin{aligned} p(\Delta | \beta, e, q, D) &\propto \prod_{j=1}^J \prod_{k=1}^j \left\{ \frac{\delta_k^{q_{jk}}}{\left(\sum_{i=1}^j \delta_i \right)^{q_{jk}}} \right\} \left\{ \prod_{j=1}^J \left[(s_j - s_{j-1}) \sum_{i=1}^j \delta_i \right]^{d_j} \right\} \\ &\quad \times \left\{ \exp \left\{ - \sum_{j=1}^J \sum_{k=1}^{d_j} e_{jk} T_j(\Delta) u_{jk}(\beta) \right\} \right\} \left\{ \prod_{j=1}^J \exp \{-\delta_j(a_j + b_j)\} \right\} \\ &\quad \times \left\{ \prod_{j=1}^J \pi(\delta_j) \right\}. \end{aligned} \tag{7.11}$$

Equation (7.11) can be simplified further. Define $w_{jk}(\beta) = u_{jk}(\beta)e_{jk}$ and let $w_{j+}(\beta) = \sum_{k=1}^{d_j} w_{jk}(\beta)$.

Since

$$\exp \left\{ - \sum_{j=1}^J \sum_{i=1}^j \delta_i w_{j+}(\beta) (s_j - s_{j-1}) \right\} = \prod_{j=1}^J \exp \left\{ - \delta_j \sum_{k=j}^J w_{k+}(\beta) (s_k - s_{k-1}) \right\},$$

we can write (7.11) as

$$\begin{aligned} p(\Delta | \beta, e, q, D) &\propto \left\{ \prod_{j=1}^J \delta_j^{\sum_{k=j}^J q_{kj}} \right\} \\ &\quad \times \left\{ \prod_{j=1}^J \exp \left\{ - \delta_j \left(a_j + b_j + \sum_{k=j}^J w_{k+}(\beta) (s_k - s_{k-1}) \right) \right\} \right\} \\ &\quad \times \left\{ \prod_{j=1}^J \pi(\delta_j) \right\}. \end{aligned} \tag{7.12}$$

We see that from (7.12) that given the latent variables (e, q) , the posterior density of Δ consists of a product of the marginal posterior densities of the δ_j 's, thus implying independence. We use (7.12) in the Gibbs sampling scheme to sample Δ .

The posterior density of $\beta|\Delta, D$ is given by

$$\begin{aligned} p(\beta|\Delta, D) &\propto \left\{ \prod_{j=1}^J \exp \{-\delta_j(a_j + b_j)\} \right\} \\ &\quad \times \left\{ \prod_{j=1}^J \prod_{k=1}^{d_j} (1 - \exp \{-u_{jk}(\beta)T_j(\Delta)\}) \right\} \pi(\beta). \end{aligned} \quad (7.13)$$

Therefore, to obtain samples from $[\beta, \Delta | D]$, we use a Gibbs sampling scheme to sample from the following four distributions:

- a) $[\Delta | \beta, e, q, D]$
- b) $[e | \beta, \Delta, q, D]$
- c) $[q | \beta, \Delta, e, D]$
- d) $[\beta | \Delta, D]$.

We use a rejection algorithm with the gamma density as the envelope to obtain a sample from the distribution in a). The gamma density used in the rejection algorithm is proportional to

$$\left\{ \prod_{j=1}^J \delta_j^{\sum_{k=j}^J q_{kj}} \right\} \left\{ \prod_{j=1}^J \exp \left\{ -\delta_j \left(a_j + b_j + \sum_{k=j}^J w_{k+}(\beta)(s_k - s_{k-1}) \right) \right\} \right\}, \quad (7.14)$$

which has mode equal to

$$\tilde{m} = \frac{\sum_{k=j}^J q_{kj}}{a_j + b_j + \sum_{k=j}^J w_{k+}(\beta)(s_k - s_{k-1})}.$$

Now, we use the gamma envelope in (7.14) along with the mode \tilde{m} in a standard rejection algorithm to decide upon acceptance or rejection of a sample from $[\Delta|\beta, e, q, D]$.

Distributions b) and c) are quite straightforward to sample from since they correspond to truncated exponential and multinomial distributions, respectively. Specifically,

$$\begin{aligned} p(e|\beta, \Delta, q, D) &\propto \exp \left\{ - \sum_{j=1}^J \sum_{k=1}^{d_j} e_{jk} u_{jk}(\beta) T_j(\Delta) \right\} \\ &\quad \times \left\{ \prod_{j=1}^J \prod_{k=1}^{d_j} I(e_{jk} \leq 1) \right\}, \end{aligned}$$

where $I(\cdot)$ is the indicator function and

$$p(q|\beta, \Delta, e, D) \propto \prod_{j=1}^J \delta_j^{\sum_{k=1}^J q_{kj}}.$$

Cycling through a), b), and c) via Gibbs will yield samples from $[\Delta|\beta, D]$.

Once a sample of Δ is obtained from $[\Delta|\beta, D]$, we complete the Gibbs cycle by sampling from $[\beta|\Delta, D]$. To obtain a sample β from this distribution, we first observe that $[\beta|\Delta, D]$ is log-concave in each component of β .

Therefore, we may directly use the algorithm of Gilks and Wild (1992) to sample from this posterior distribution.

To show that $[\beta|\Delta, D]$ is log-concave in each component of β , it suffices to show that

$$\frac{\partial^2 \log p(\beta|\Delta, D)}{\partial \beta_r^2} \leq 0$$

for all $r = 1, \dots, p$.

Letting

$$\begin{aligned} A_{jk}(\beta, \Delta) &= u_{jk}(\beta) T_j(\Delta), \\ B_{jk}(\beta, \Delta) &= 1 - \exp \{-A_{jk}(\beta, \Delta)\}, \\ C_{jk}(\beta, \Delta) &= 1 - A_{jk}(\beta, \Delta) - \exp \{-A_{jk}(\beta, \Delta)\}, \end{aligned}$$

we get

$$\frac{\partial^2 \log(p(\beta|\Delta, D))}{\partial \beta_r^2} = \sum_{j=1}^J \sum_{k=1}^{d_j} \left\{ x_{jkr}^2 A_{jk}(\beta, \Delta) B_{jk}^{-2}(\beta, \Delta) \exp \{-A_{jk}(\beta, \Delta)\} C_{jk}(\beta, \Delta) \right\} \quad (7.15)$$

$$- \sum_{j=1}^J \sum_{i=j+1}^J \sum_{k=1}^{d_i} \left\{ \delta_j x_{ikr}^2 \exp \{x'_{ik}\beta\} (s_{i-1} - s_{j-1}) \right\} \quad (7.16)$$

$$- \sum_{j=1}^J \sum_{i=j}^J \sum_{k=d_i+1}^{d_i+c_i} \left\{ \delta_j x_{ikr}^2 \exp \{x'_{ik}\beta\} (s_i - s_{j-1}) \right\} + \frac{\partial^2 \log \pi(\beta)}{\partial \beta_r^2}. \quad (7.17)$$

We first note that since we are using a normal prior for β , it is well known that $\frac{\partial^2 \log \pi(\beta)}{\partial \beta_r^2} < 0$.

Second, we clearly see that (7.16) and (7.17) are negative. Thus, to show that $\frac{\partial^2 \log(p(\beta|\Delta, D))}{\partial \beta_r^2} \leq 0$, it is enough to show that (7.15) is negative.

It suffices to show that $C_{jk}(\beta, \Delta)$ in (7.15) is negative, since all of the other terms in the summand of (7.15) are positive. We see that $C_{jk}(\beta, \Delta)$ is of the form $f(x) = 1 - \exp\{x\} - \exp\{-\exp\{x\}\}$. Clearly, when $x > 0$, $f(x) < 0$. For $x \in (-\infty, 0)$, we see that $f(x)$ is a monotonic decreasing function and $\lim_{x \rightarrow -\infty} f(x) = 0$. Thus $f(x) < 0$ for all $x \in R^1$, and thus $C_{ik}(\beta, \Delta) \leq 0$ for all (β, Δ) .

Thus $[\beta|\Delta, D]$ is log-concave in each component of β .

Bayesian Inference Using the Partial Likelihood

The partial likelihood itself has a Bayesian motivation as discussed in Kalbfleisch (1978, JRSS-B) and extended more recently in Sinha, Ibrahim, and Chen (2003, Biometrika).

For Bayesian inference, it is valid to use the partial likelihood as your “likelihood” and this does away with discretizing the time axis and specifying priors on the hazard altogether.

In this framework, we only have β in the model, and the posterior distribution of β is given by

$$p(\beta|D) \propto \text{PL}(\beta) \pi(\beta)$$

where $\pi(\beta)$ can be taken to be a normal distribution.

Frailty Models

In studies of survival, the hazard function for each individual may depend on a set of risk factors or explanatory variables, but usually not all such variables are known or measurable.

This unknown and unobservable risk factor of the hazard function is often termed as the individual's heterogeneity or *frailty*.

Frailty models are becoming increasing popular in multivariate survival analysis since they allow us to model the association between the individual survival times within subgroups or *clusters* of subjects.

Proportional Hazards Model with Frailty

The most common type of frailty model is called the shared-frailty model, which is an extension of the Cox proportional hazards model.

This model can be derived as follows: Let y_{ij} denote the survival time for the j^{th} subject in the i^{th} cluster, $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, m_i$.

Thus, m_i represents the number of subjects in the i^{th} cluster, and therefore we have a total of $N = \sum_{i=1}^n m_i$ subjects.

In the shared frailty model, we assume that the conditional hazard function of y_{ij} given the unobserved frailty random variable w_i for the i^{th} cluster and the fixed covariate vector x_{ij} is given by

$$h(y|w_i, x_{ij}) = h_0(y)w_i \exp\{x'_{ij}\beta\}, \quad (7.18)$$

$$i = 1, 2, \dots, n, j = 1, 2, \dots, m_i.$$

Here, β is the $p \times 1$ vector of unknown regression coefficients, $h_0(\cdot)$ is an unknown baseline hazard function common to every subject, and x_{ij} is the $p \times 1$ covariate vector for the j^{th} subject in the i^{th} cluster (may be time dependent).

Another formulation is to take

$$h(y|w_i, x_{ij}) = h_0(y) \exp\{w_i + x'_{ij}\beta\},$$

It is common method to use a parametric distribution for the frailty w_i .

For (7.18), finite mean frailty distributions such as the gamma and log-normal are very popular in the literature in spite of their theoretical limitations.

Other alternatives include using infinite mean distributions such as the positive stable distribution.

The gamma distribution for (7.18) is the most commonly used finite mean distribution to model the frailty term w_i .

For finite mean frailty distributions, we need the mean of the frailty distribution to be unity in order for the parameters of the model to be identifiable.

Thus, for the gamma frailty model, the w_i 's are assumed to be i.i.d. with

$$w_i \sim \text{gamma}(\kappa^{-1}, \kappa^{-1}),$$

where κ is the (unknown) variance of the w_i 's.

Thus, larger values of κ imply greater heterogeneity among clusters.

Letting $w = (w_1, w_2, \dots, w_n)'$, we have

$$\pi(w) \propto \prod_{i=1}^n w_i^{\kappa^{-1}-1} \exp\{-\kappa^{-1}w_i\}. \quad (7.19)$$

Bayesian methods are attractive for these models since they easily allow an analysis using the full likelihood and inference does not rely on asymptotics.

Weibull Model with Gamma Frailties

Let ν_{ij} denote the censoring indicator variable, taking value 1 if the j^{th} subject ($j = 1, 2, \dots, m_i$) of the i^{th} cluster ($i = 1, 2, \dots, n$) fails and 0 otherwise. Hence, y_{ij} is a failure time if $\nu_{ij} = 1$ and a censoring time otherwise.

Further, let $\nu = (\nu_{11}, \nu_{12}, \dots, \nu_{nm_n})'$, $y = (y_{11}, y_{12}, \dots, y_{nm_n})'$, and $X = (X_1, X_2, \dots, X_n)$, where X_i is the $m \times p$ matrix of covariates for the i^{th} cluster.

Let $D = (X, \nu, y, w)$ denote the complete data and let $D_{obs} = (X, \nu, y)$ denote the observed data.

Let the Weibull baseline hazard function be given by

$$h_0(y_{ij}) = \gamma\alpha y_{ij}^{\alpha-1},$$

where (γ, α) are the parameters of the Weibull distribution.

The hazard function is given by

$$h(y_{ij}|x_{ij}, w_i) = \gamma\alpha w_i y_{ij}^{\alpha-1} \theta_{ij}, \quad (7.20)$$

where $\theta_{ij} = \exp\{x'_{ij}\beta\}$.

The complete data likelihood is given by

$$L(\beta, \gamma, \alpha) = \prod_{i=1}^n \prod_{j=1}^{m_i} (\gamma \alpha y_{ij}^{\alpha-1} w_i \theta_{ij})^{\nu_{ij}} \exp \{-\gamma y_{ij}^\alpha \theta_{ij} w_i\}. \quad (7.21)$$

The likelihood function of (β, γ, α) based on the observed data D_{obs} can be obtained by integrating out the w_i 's from (7.21) with respect to the density $\pi(w)$ given in (7.19).

The observed data likelihood is far too complicated to work with, and thus it is difficult to evaluate the joint posterior distribution of (β, γ, α) analytically.

To circumvent this problem, we use the Gibbs sampler to generate samples from the joint posterior distribution.

Let $\pi(\cdot)$ denote the prior density of its argument and let

$$S = \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij}^\alpha \theta_{ij} w_i.$$

The full conditional distribution of each w_i is a gamma distribution, i.e.,

$$(w_i|\beta, \alpha, \gamma, D_{obs}) \sim \text{gamma} \left\{ \kappa^{-1} + \sum_{j=1}^{m_i} \nu_{ij}, \kappa^{-1} + \gamma \sum_{j=1}^{m_i} y_{ij}^\alpha \theta_{ij} \right\}, \quad (7.22)$$

for $i = 1, 2, \dots, n$.

Letting $\eta = \kappa^{-1}$, the full conditional distribution of η is given by

$$p(\eta|\beta, w, D_{obs}) \propto \prod_{i=1}^n w_i^{\eta-1} \eta^{n\eta} \frac{\exp\{-\eta \sum_{i=1}^n w_i\}}{[\Gamma(\eta)]^n} \pi(\eta). \quad (7.23)$$

The full conditional of β is given by

$$p(\beta|\eta, w, D_{obs}) \propto \exp \left\{ \beta' \sum_{i=1}^n \sum_{j=1}^{m_i} \nu_{ij} x_{ij} - \gamma S \right\} \pi(\beta). \quad (7.24)$$

If a priori $\gamma \sim \text{gamma}(\rho_1, \rho_2)$, then the full conditional of γ is a gamma distribution given by

$$(\gamma|\beta, \alpha, w, D_{obs}) \sim \text{gamma} \left\{ \rho_1 + \sum_{i=1}^n \sum_{j=1}^{m_i} \nu_{ij}, \rho_2 + S \right\}. \quad (7.25)$$

Finally, the full conditional of α is given by

$$p(\alpha|\beta, \gamma, w, D_{obs}) \propto \left(\prod_{i=1}^n \prod_{j=1}^{m_i} y_{ij}^{\nu_{ij}} \right)^{\alpha-1} \alpha^{\sum_{i=1}^n \sum_{j=1}^{m_i} \nu_{ij}} \exp\{-\gamma S\} \pi(\alpha). \quad (7.26)$$

A priori, it is common to take $\alpha \sim \text{gamma}(a_1, a_2)$, so that

$$\pi(\alpha) \propto \alpha^{a_1-1} \exp\{-\alpha a_2\}.$$

With these choices of priors, it can be shown that each of the above full conditional densities in (7.22)–(7.26) is log-concave.

The frailty model in (7.20) is a multiplicative frailty model.

An alternative modeling strategy is an additive frailty model,

$$h(y_{ij}|x_{ij}, b_i) = \xi_{ij} \alpha y_{ij}^{\alpha-1},$$

$$\log(\xi_{ij}) = \zeta + x'_{ij}\beta + b_i$$

The b_i 's are assumed i.i.d. $N(0, \kappa^{-1})$, and κ is given a gamma(ϕ_1, ϕ_2) prior.

The prior for α is gamma(a_1, a_2), and ζ is given a normal prior.

It is expected that both the multiplicative and additive hazard Weibull model formulations will yield similar inferences since this additive frailty model is actually a multiplicative frailty model with a log-normal frailty distribution.

Example 7.10: Kidney Data (Frailty Model)

We consider the times of infection from the time of insertion of the catheter for 38 kidney patients using portable dialysis equipment.

The first two columns give the time of first and second infection. The third and fourth columns give the failure indicators for the first and second infections, respectively.

Occurrence of infection is indicated by 1, and 0 means that the infection time is right censored.

Other columns describe age at the time of each infection, sex (0 = male, 1 = female), and the disease (0 = other, 1 = GN, 2 = AN, 3 = PKD).

Data in *kidney.dat*

time1	time2	infection1	infection2	age1	age2	sex	disease
8	16	1	1	28	28	0	0
23	13	1	0	48	48	1	1
22	28	1	1	32	32	0	0
447	318	1	1	31	32	1	0
...
54	16	0	0	42	42	1	0
6	78	0	1	52	52	1	3
63	8	1	0	60	60	0	3

We consider five covariates in the model: age_{ij} , sex_i , disease_{i1} , disease_{i2} , and disease_{i3} , where disease_{ij} is an indicator for disease j . For ease of exposition, we label these covariates as $(x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5})$, respectively.

The model is thus given by

$$y_{ij} \sim \text{Weibull}(\alpha, \gamma_{ij}),$$

where $i = 1, 2, \dots, 38$, $j = 1, 2$, and

$$\begin{aligned}\log(\gamma_{ij}) &= \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} \\ &\quad + \beta_4 x_{ij4} + \beta_5 x_{ij5} + b_i,\end{aligned}$$

where $b_i \sim N(0, \sigma^2)$. Let $\tau = \frac{1}{\sigma^2}$.

We take $\beta \sim N_6(0, 10^4 I_6)$ and $\tau \sim \text{gamma}(0.001, 0.001)$. Further, we take $\alpha \sim \text{gamma}(1, 1)$.

We fit the model in both JAGS and SAS using PROC MCMC.

We write out the model in the file *kidneyJAGS.txt*, as shown below:

```
model{  
  for(i in 1:N){  
    for(j in 1:M){  
      # Survival and censoring times  
      is.censored[i,j] ~ dinterval(y[i,j], y.cen[i,j])  
      y[i,j] ~ dweib(alpha, mu[i,j])  
      log(mu[i,j]) <- beta0 + beta.age * age[i,j] + beta.sex * sex[i] + beta.d1 * d1[i] +  
                      beta.d2 * d2[i] + beta.d3 * d3[i] + b[i]  
    }  
    # Random effects  
    # Specify precision (not variance or standard deviation) in normal distribution  
    b[i] ~ dnorm(0, tau)  
  }  
  
  # Prior distributions  
  beta0 ~ dnorm(0, 0.0001)  
  beta.age ~ dnorm(0, 0.0001)  
  beta.sex ~ dnorm(0, 0.0001)  
  beta.d1 ~ dnorm(0, 0.0001)  
  beta.d2 ~ dnorm(0, 0.0001)  
  beta.d3 ~ dnorm(0, 0.0001)  
  # Use rate parameterization for gamma distribution  
  alpha ~ dgamma(1, 1)  
  tau ~ dgamma(0.001, 0.001)  
  
  sigma2 <- 1/tau  
  sigma <- sqrt(sigma2)  
}
```

R code:

```
# Prepare input data for JAGS model - create y, y.cen, is.censored variables
kidney <- read.table( "kidney.dat", header = TRUE )
y1 <- ifelse( kidney[,3] == 1, kidney[,1], NA )
y2 <- ifelse( kidney[,4] == 1, kidney[,2], NA )
y <- cbind(y1, y2)
y.cen1 <- ifelse( is.na(y1), kidney[,1], max(kidney[,1:2]) )
y.cen2 <- ifelse( is.na(y2), kidney[,2], max(kidney[,1:2]) )
y.cen <- cbind(y.cen1, y.cen2)
is.censored <- 1 - kidney[,3:4]
data.jags <- list( N = nrow(kidney), M = 2, y = y, y.cen = y.cen,
                   is.censored = is.censored, age = kidney[,5:6], sex = kidney$sex,
                   d1 = ifelse( kidney$disease == 1, 1, 0),
                   d2 = ifelse( kidney$disease == 2, 1, 0),
                   d3 = ifelse( kidney$disease == 3, 1, 0) )

# Initialize missing values of y
y.inits <- ifelse( is.na(y), y.cen + 1, NA )
jags.inits <- list( list(y = y.inits) )

# Set up model using JAGS ("rjags" package)
library(rjags)
jags.mod <- jags.model( file = "kidneyJAGS.txt", data = data.jags, inits = jags.inits,
                        n.adapt = 5000, n.chains = 1 )

# Specify names of parameters to sample
jags.vars <- c( "beta0", "beta.age", "beta.sex", "beta.d1", "beta.d2", "beta.d3", "alpha", "sigma2" )
jags.samps <- coda.samples( jags.mod, n.iter = 50000, thin = 5, variable.names = jags.vars )

# Summary of parameter estimates
summary( jags.samps )
```

SAS code:

```
/* Read in data */
data kidney;
  infile "kidneySAS.dat" firstobs=2;
  input id infection time fail age sex disease;
run;

/* Create new version of data that recodes disease as three indicator variables */
data kidney2;
  set kidney;
  if disease = 1 then dis1 = 1; else dis1 = 0;
  if disease = 2 then dis2 = 1; else dis2 = 0;
  if disease = 3 then dis3 = 1; else dis3 = 0;
run;
```

SAS code (cont.):

```
/* Frailty model */
title 'Frailty Model';
ods graphics on;
proc mcmc data=kidney2 outpost=outFrailty nbi=10000 nmc=100000 thin=10 seed=779 maxtune=50;
    parms alpha 1;
    prior alpha ~ gamma(1, iscale=1);
    array beta[6];
    parms beta1-beta6 0;
    prior beta: ~ normal(0, var=10000);
    array re[38];
    parms re1-re7;
    parms re8-re14;
    parms re15-re21;
    parms re22-re28;
    parms re29-re35;
    parms re36-re38;
    parms sig2 .1;
    prior re: ~ normal(0, v=sig2);
    prior sig2 ~ igamma(0.001, scale=0.001);
    lambda = beta1 + beta2*age + beta3*sex + beta4*dis1 +
        beta5*dis2 + beta6*dis3 + re[id];
    loglike = fail*(log(alpha) + (alpha-1)*log(time) + lambda)
        exp(lambda)*(time**alpha);
    model general(loglike);
run;
ods graphics off;
```

Analysis in JAGS and SAS

The table below gives posterior estimates of the regression parameters from JAGS based on 50,000 iterations (thin by 5) after a 5000 iteration burn-in.

The SAS results based on 100,000 iterations (thin by 10) after a 10,000 iteration burn-in are also shown.

Parameter	JAGS Estimate (SD)	SAS Estimate (SD)
β_0	-4.625 (0.935)	-4.773 (1.022)
β_1 (age)	0.004 (0.015)	0.004 (0.017)
β_2 (sex)	-1.948 (0.521)	-1.990 (0.543)
β_3 (disease ₁)	0.105 (0.559)	0.138 (0.593)
β_4 (disease ₂)	0.622 (0.559)	0.670 (0.609)
β_5 (disease ₃)	-1.192 (0.828)	-1.158 (0.899)
α	1.214 (0.171)	1.250 (0.177)
σ^2	0.561 (0.521)	0.688 (0.582)

We can extend the piecewise exponential models and partial likelihood models discussed earlier to the frailty setting.

We just replace λ_j by $w_i \lambda_j$ in the piecewise exponential model to get the multiplicative model, and specify a gamma distribution for w_i .

To get the additive model we replace λ_j by $\exp\{w_i\} \lambda_j$, and specify a normal distribution for w_i .

In the partial likelihood model we replace $x_i' \beta$ with $x_i' \beta + b_i$.

Cure Rate Models

The cure rate model has been used for modeling time-to-event data for various types of diseases where a significant proportion of patients are “cured”.

Cure rate models are appropriate when a plateau occurs in the survival function after sufficient follow-up.

To demonstrate such a phenomenon, we consider a recent phase III clinical trial in malignant melanoma (E1684) undertaken by the Eastern Cooperative Oncology Group (ECOG). The graph in Figure 7.4 gives the Kaplan-Meier survival curve for $n = 285$ patients in E1684, with the survival time given in years.

We see from Figure 7.4 that a plateau in the curve occurs at approximately 0.30, suggesting that 30% fraction of patients are “cured” after sufficient follow-up.

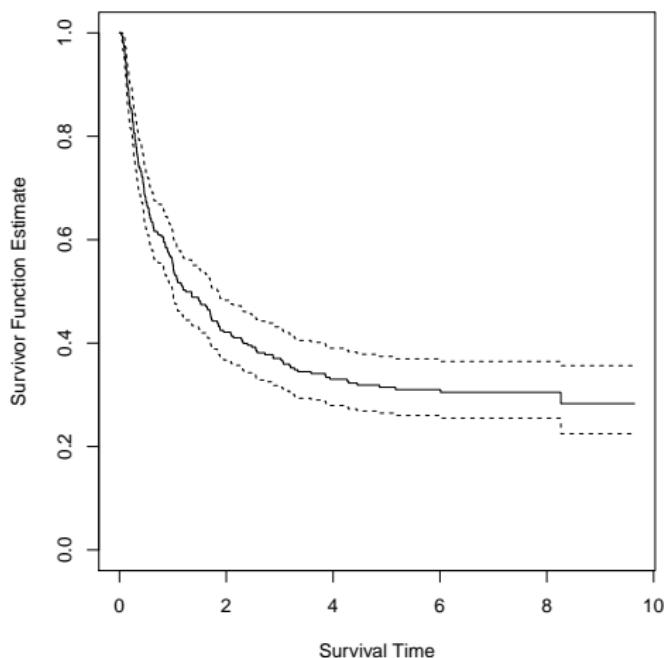


Figure 7.4: Kaplan-Meier plot for E1684 data.

Development of the Cure Rate Model

To describe the cure rate model, consider a study where N is the number of carcinogenic cells for an individual, which is assumed to have a Poisson distribution with mean θ .

Let Z_i be the random time for the i^{th} carcinogenic tumor cell to produce a detectable tumor, which can also be viewed as an incubation time or promotion time.

N and Z_i are latent variables that are not observed.

Assume that Z_i , $i = 1, 2, \dots$ are independent and identically distributed with a common distribution function $F(y) = 1 - S(y)$ and independent of N .

Let $Y = \min \{Z_i, 0 \leq i \leq N\}$, the tumor latency time. Y is the observed time to event, such as time-to-relapse in cancer.

The Survival Function

$$\begin{aligned} S_p(y) &= P(\text{no cancer by time } y) \\ &= P(N = 0) + P(Z_1 > y, \dots, Z_N > y, N \geq 1) \\ &= \exp\{-\theta\} + \sum_{k=1}^{\infty} S(y)^k \frac{\theta^k}{k!} \exp\{-\theta\} \\ &= \exp\{-\theta + \theta S(y)\} \\ &= \exp\{-\theta F(y)\}. \end{aligned}$$

The cure fraction (i.e., cure rate) is given by

$$P(N = 0) = \exp\{-\theta\} = \lim_{y \rightarrow \infty} S_p(y).$$

Note that $S_p(y)$ is not a proper survival function as

$$\lim_{y \rightarrow \infty} S_p(y) = \exp\{-\theta\} \neq 0.$$

We specify a parametric form for $F(\cdot)$, such as a Weibull or gamma distribution.

We denote the indexing parameter (possibly vector valued) by ψ , and thus write $F(\cdot|\psi)$ and $S(\cdot|\psi)$.

For example, if $F(\cdot|\psi)$ corresponds to a Weibull distribution, then $\psi = (\alpha, \lambda)'$, where α is the shape parameter and λ is the scale parameter.

The Likelihood Function

To set up the likelihood function, we first define the following:

- ▶ y_i : survival time (failure time or right-censored)
- ▶ ν_i : the censoring indicator; $\nu_i = 1$ if y_i is the failure time and 0 if y_i is right censored.
- ▶ N_i : the number of carcinogenic cells
- ▶ $x'_i = (x_{i1}, \dots, x_{ip})$: the $p \times 1$ vector of covariates
- ▶ n : the sample size
- ▶ $\beta = (\beta_1, \dots, \beta_p)'$: the vector of regression coefficients

We model $\theta_i = \exp\{x'_i \beta\}$.

We denote the observed data as $D_{obs} = (n, y, \nu)$, and we denote the complete data as $D = (n, y, \nu, N)$, where $y = (y_1, y_2, \dots, y_n)'$, $\nu = (\nu_1, \nu_2, \dots, \nu_n)'$, and $N = (N_1, N_2, \dots, N_n)'$.

The complete data likelihood function is given by

$$\begin{aligned} L(\beta, \psi | D) &= \left(\prod_{i=1}^n S(y_i | \psi)^{N_i - \nu_i} (N_i f(y_i | \psi))^{\nu_i} \right) \\ &\quad \times \exp \left\{ \sum_{i=1}^n [N_i x'_i \beta - \log(N_i!) - \exp\{x'_i \beta\}] \right\}. \end{aligned}$$

We assume a Weibull density for $f(y_i|\psi)$, so that

$$f(y|\psi) = \alpha y^{\alpha-1} \exp\{\lambda - y^\alpha \exp\{\lambda\}\}.$$

After summing out N , we obtain

$$L(\beta, \psi | D_{obs}) = \prod_{i=1}^n \left(\theta_i f(y_i|\psi) \right)^{\nu_i} \exp\{-\theta_i(1 - S(y_i|\psi))\}, \quad (7.27)$$

where $\theta_i = \exp\{x'_i \beta\}$.

We take

$$\begin{aligned}\pi(\beta, \psi) &\propto \pi(\psi), \\ \pi(\psi) &= \pi(\alpha)\pi(\lambda), \\ \pi(\lambda) &= N(\mu_0, \sigma_0^2), \\ \pi(\alpha) &\propto \alpha^{\delta_0 - 1} \exp\{-\tau_0\alpha\},\end{aligned}$$

where δ_0 and τ_0 are two specified hyperparameters.

With these specifications, the posterior distribution of (β, ψ) based on the observed data $D_{obs} = (n, y, X, \nu)$ is given by

$$p(\beta, \psi | D_{obs}) \propto L(\beta, \psi | D_{obs})\pi(\alpha)\pi(\lambda).$$

We can also specify the power prior for this model as

$$\pi(\beta, \psi | D_0) \propto L(\beta, \psi | D_{0,obs})^{a_0} \pi_0(\beta, \psi).$$

where $D_{0,obs} = (n_0, y_0, X_0, \nu_0)$.

Example 7.11: Melanoma Data (Cure Rate Model)

We now consider the E1684 melanoma data using a cure rate model. We use the likelihood specified in (7.27), and we fit the model using both JAGS and PROC MCMC in SAS.

Note that JAGS and SAS use different parameterizations of the Weibull distribution. In JAGS, the Weibull density is

$$f(y) = \alpha\gamma y^{\alpha-1} \exp\{-\gamma y^\alpha\}, \quad \alpha > 0, \gamma > 0,$$

denoted by $\text{Weibull}(\alpha, \gamma)$.

In SAS, the Weibull density is

$$f(y) = \frac{\alpha}{\eta} \left(\frac{y}{\eta}\right)^{\alpha-1} \exp\left\{-\left(\frac{y}{\eta}\right)^\alpha\right\}, \quad \alpha > 0, \eta > 0,$$

denoted by $\text{Weibull}(\alpha, \eta)$.

We consider treatment as a single covariate with β_1 as the corresponding effect. Let $\beta = (\beta_0, \beta_1)'$.

Let $\lambda = \log(\gamma) = -\alpha \log(\eta)$. A priori, we take

$$\begin{aligned}\beta &\sim N_2(0, 10^4 I_2), \\ \alpha &\sim \text{gamma}(2, 2), \\ \lambda &\sim N(0, 10).\end{aligned}$$

We write out the model in the file *e1684JAGScr.txt*, as shown below:

```
model{  
  c <- 10000  
  # Likelihood part of Bayesian inference  
  for(i in 1:n){  
    s[i] <- exp( -exp(lambda) * y[i]^alpha )  
    xb[i] <- beta0 + betai * trt[i]  
    theta[i] <- exp(xb[i])  
    L1[i] <- -theta[i] * (1 - s[i])  
    L2[i] <- nu[i] * ( xb[i] + log(alpha) + lambda +  
      (alpha - 1) * log(y[i]) - exp(lambda) * y[i]^alpha )  
    mu[i] <- -(L1[i] + L2[i]) + c  
    zeros[i] ~ dpois(mu[i])  
  }  
  
  # Priors  
  beta0 ~ dnorm(0, 0.0001)  
  betai ~ dnorm(0, 0.0001)  
  lambda ~ dnorm(0, 0.1)  
  alpha ~ dgamma(2, 2)  
}
```

R code:

```
# Prepare input data for JAGS model
e1684.90 <- read.table( "mina-e1684-e1690.txt", header = TRUE )
e1684 <- e1684.90[ e1684.90$study == 1684, c(10,11,3:6) ]
data.jags <- list( y = e1684$faultime,
                   nu = e1684$rfscens,
                   trt = e1684$trt,
                   n = nrow(e1684),
                   zeros = numeric(nrow(e1684)) )

# Set up model using JAGS ("rjags" package)
library(rjags)
start.time <- Sys.time()
jags.mod <- jags.model( file = "e1684JAGScr.txt", data = data.jags,
                        n.adapt = 1000, n.chains = 1 )

# Specify names of parameters to sample
jags.vars <- c( "beta0", "beta1", "alpha", "lambda" )
jags.samps <- coda.samples( jags.mod, n.iter = 10000, thin = 1, variable.names = jags.vars )
end.time <- Sys.time()
end.time - start.time

# Summary of parameter estimates
summary( jags.samps )
```

SAS code:

```
/* Read in data */
data e1684;
  infile "e1684SAS.dat" firstobs=2;
  input failtime rfscens trt;
run;

/* Cure rate model */
proc mcmc data=e1684 outpost=cout nmc=10000 seed=779;
  parms alpha 1 lambda 1 (beta0 beta1) 1;
  prior beta: ~ normal(0, var=10000);
  prior alpha ~ gamma(2, iscale=2);
  prior lambda ~ normal(0, var=10);
  eta = exp(-lambda/alpha);
  theta = exp(beta0 + beta1 * trt);
  llike = rfscens * (log(theta) + logpdf('weibull', failtime, alpha, eta)) -
    theta * (1 - sdf('weibull', failtime, alpha, eta));
  model general(llike);
run;
```

Analysis in JAGS and SAS

The table below gives posterior estimates of the regression parameters from both JAGS and SAS, each based on 10,000 iterations after a 1000 iteration burn-in.

Posterior Summaries Using the Frailty Model for Kidney Data

Parameter	JAGS Estimate (SD)	SAS Estimate (SD)
β_0	0.387 (0.107)	0.385 (0.104)
β_1 (trt)	-0.354 (0.147)	-0.347 (0.140)
α	1.013 (0.059)	1.012 (0.059)
λ	-0.476 (0.101)	-0.474 (0.096)

Example 7.12: Melanoma Data (Cure Rate with Power Prior)

We discuss an analysis by Chen, Ibrahim, and Sinha (1999, JASA) for the E1684 trial using the power prior ($n = 285$). Data from a previous trial, E1673 ($n_0 = 650$), were used as the historical data.

The results are shown in Table 7.2.

Table 7.2: Posterior Estimates of the Model Parameters

a_0	Variable	Posterior	Posterior	95% HPD
		Mean	Std Dev	Interval
0	intercept	0.09	0.11	(-0.12, 0.30)
	age	0.09	0.07	(-0.05, 0.23)
	gender	-0.12	0.16	(-0.44, 0.19)
	ps	-0.23	0.26	(-0.73, 0.28)
	α	1.31	0.09	(1.15, 1.48)
	λ	-1.36	0.12	(-1.60, -1.11)
0.14	intercept	0.25	0.10	(0.05, 0.45)
	age	0.12	0.06	(-0.00, 0.24)
	gender	-0.20	0.14	(-0.47, 0.07)
	ps	-0.09	0.22	(-0.53, 0.31)
	α	1.06	0.06	(0.95, 1.17)
	λ	-1.62	0.12	(-1.85, -1.39)
0.29	intercept	0.26	0.09	(0.08, 0.43)
	age	0.13	0.06	(0.02, 0.24)
	gender	-0.24	0.12	(-0.48, 0.00)
	ps	-0.01	0.19	(-0.38, 0.35)
	α	1.03	0.05	(0.93, 1.13)
	λ	-1.70	0.11	(-1.91, -1.50)
1	intercept	0.22	0.06	(0.11, 0.35)
	age	0.16	0.04	(0.08, 0.24)
	gender	-0.32	0.09	(-0.50, -0.15)
	ps	0.14	0.13	(-0.11, 0.39)
	α	1.00	0.04	(0.93, 1.07)
	λ	-1.82	0.08	(-1.97, -1.67)

Table 7.2 indicates a fairly robust pattern of behavior. The estimates of the posterior mean, standard deviation, or highest posterior density (HPD) intervals of β do not change a great deal if a low or moderate weight is given to the historical data.

However, if a higher than moderate weight is given to the historical data, these posterior summaries can change a lot.

For example, when the posterior mean of a_0 is less than 0.14, we see that all of the HPD intervals for β include 0, and when the posterior mean of a_0 is greater than or equal to 0.14, some HPD intervals for β do not include 0.

The HPD interval for age does not include 0 when the posterior mean of a_0 is 0.29, and it includes 0 when less weight is given to the historical data.

This finding is interesting, since it indicates that age is a potentially important prognostic factor for predicting survival in melanoma.

Such a conclusion is not possible based on a frequentist or Bayesian analysis of the current data alone.

Residuals in Survival Analysis

In general, the Cox-Snell residual is defined as

$$r_{C_i} = \exp\{x_i' \hat{\beta}\} \hat{\Lambda}_0(y_i),$$

and recall that $-\log(\hat{S}_0(y_i)) = \hat{\Lambda}_0(y_i)$. The **Martingale Residual** for the i^{th} subject is

$$r_{M_i} = \nu_i - r_{C_i},$$

where ν_i = censoring indicator for i^{th} subject.

The **Deviance Residual** is given by

$$r_{D_i} = \text{sign}(r_{M_i}) [-2\{r_{M_i} + \nu_i \log(\nu_i - r_{M_i})\}]^{1/2}.$$

For the Weibull model,

$$\begin{aligned} h_0(y) &= \alpha y^{\alpha-1}, \\ S_0(y) &= \exp\{-y^\alpha\}. \end{aligned}$$

With covariates,

$$S_i(y) = \exp\{-\exp\{x_i' \beta\} y^\alpha\},$$

for the i^{th} subject, and the Cox-snell residual is $r_{C_i} = \exp\{x_i' \hat{\beta}\} y^{\hat{\alpha}}$.

Chapter 8:

Bayesian Methods for Clinical Trials

Bayesian Methods in the Design of Clinical Trials

In this chapter, we discuss Bayesian methods for clinical trials. See “Bayesian Approaches to Randomized Trials” by Spiegelhalter, Freedman, and Parmar (1994, JRSSA) for additional information.

Specifically, we discuss the following topics:

- 1) Sample size calculations
- 2) Interim analyses
- 3) Prior distributions
- 4) Predictions
- 5) Cancer clinical trial example
- 6) Advantages of the Bayesian approach.

Types of Clinical Trials

There are essentially three types of clinical trials: Phase I, Phase II, and Phase III.

Phase I

The Phase I clinical trial is a study intended to estimate the so-called maximum tolerable dose (MTD) of a new regimen.

In such studies, toxicity and pharmacologic information is obtained from dose escalation experiments, whereby volunteers are subjected to increasing doses of the regimen according to a predetermined schedule.

Typically, a Phase I study has 20-40 patients.

Phase II

A Phase II clinical trial is a small scale study investigating the efficacy and safety of a regimen.

Efficacy can be measured in terms of response rates and/or survival. Safety is assessed in terms of toxicity.

A typical Phase II study has 30-100 patients.

Phase III

After a regimen is shown to be reasonably effective, it is natural to compare it with the current standard treatment. This leads to a Phase III clinical trial.

A Phase III clinical trial is a randomized large scale comparative study involving two or more treatment arms.

Phase III trials can have between 250-4000 patients.

The Phase III trial is the most rigorous and extensive type of scientific clinical investigation of a new treatment.

Phase III trials are perhaps the most important of the three types of clinical trials since it involves a substantial number of patients and has the potential for having a major impact on patients lives.

Issues arising for design of Phase III studies include

- 1) Endpoints
- 2) Sample size
- 3) Randomization
- 4) Sequential monitoring
- 5) Stopping rules
- 6) Toxicity

Null Hypothesis and Ranges of Equivalence

Consider the situation in which we have two treatments A and B, and that the true treatment difference is summarized by a parameter δ , where large values of δ correspond to superiority of the new treatment.

δ may be the population mean difference between A and B or it may be the log of the hazard ratio, or it may be a difference in proportions.

We are interested in testing the hypothesis

$$H_0 : \delta = 0 \tag{8.1}$$

against a one or two-sided alternative.

In some situations, the treatments may be so unequal in their toxicity, monetary cost, etc., that it is commonly accepted that the more “costly” treatment will be required to achieve at least a certain margin of benefit, δ_i , before it can be considered.

Hence, $\delta < \delta_i$ corresponds to clinical inferiority of the new treatment.

Another value δ_s may be specified where $\delta > \delta_s$ indicates clinical superiority of the new treatment. Sometimes δ_s is referred to as the “minimal clinically worthwhile benefit.”

The interval (δ_i, δ_s) is called the **range of equivalence**. Thus if we were certain that δ is in this interval, we would be unable to make a definitive choice of treatment.

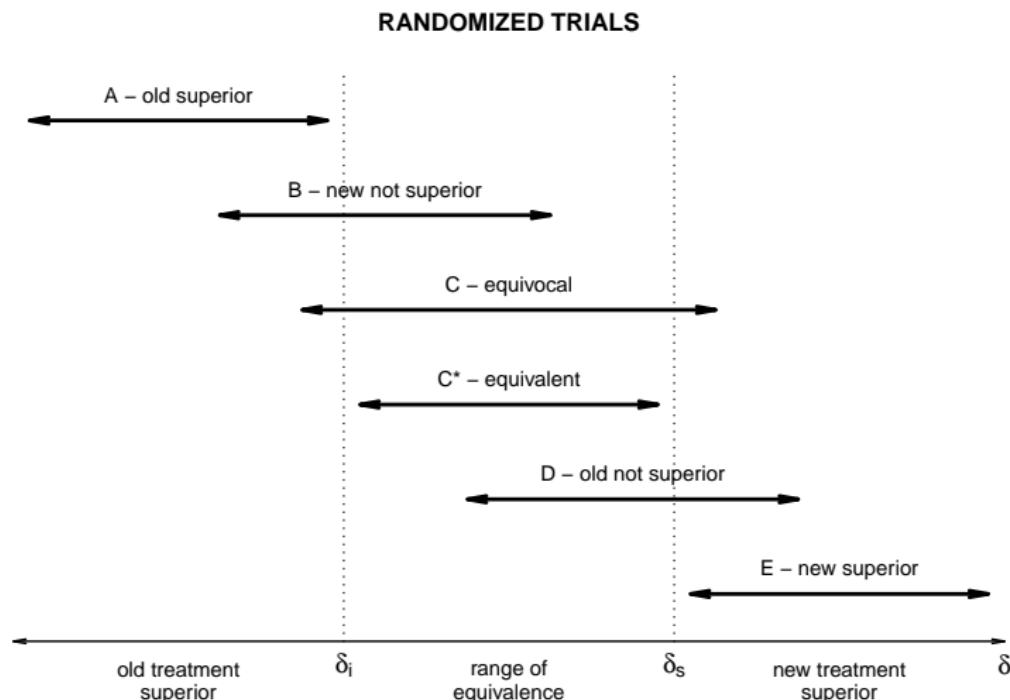


Figure 8.1: Possible situations at any point in a trial's progress, derived from superimposing an interval estimate (say 95%) on the range of equivalence.

The Bayesian Paradigm

Inference is based on the posterior distribution of the parameters. In the case of design of phase III studies, the posterior distribution as well as the preposterior marginal distribution of the data will be relevant in determining sample size and for sequential monitoring.

Bayesians believe in the likelihood principle. Two different sampling plans that result in the same likelihood will lead to the same inference about the unknown parameters. Frequentists are tied down to a design.

Bayes theorem is

$$p(\delta | x) = \frac{p(x | \delta) p(\delta)}{\int_{\Delta} p(x | \delta) p(\delta) d\delta}.$$

The quantity

$$p(x) = \int_{\Delta} p(x | \delta) p(\delta) d\delta$$

is sometimes called the preposterior marginal distribution of x .

Thus,

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Written in terms of a hypothesis H , we have

$$p(H | \text{data}) \propto p(\text{data} | H) p(H),$$

where $p(H)$ is the prior probability of H .

Generally speaking, Bayesian approaches have several advantages over other methods, including the following:

- 1) Incorporation of real prior information to improve (or change) a design or analysis.
- 2) Interpretation (e.g., confidence intervals)
- 3) Computation (MCMC methods)
- 4) Exact inference (no asymptotics needed with MCMC).

Sample Size

Let δ denote the treatment difference. Suppose we wish to test

$$H_0 : \delta = 0$$

versus $H_1 : \delta \neq 0$ or $H_1 : \delta > 0$.

In the classical approach, the sample size is chosen to guarantee, under a chosen value of δ , say δ_a for the alternative hypothesis, an acceptably high power (e.g., $1 - \beta \geq 80\%$) of rejecting H_0 by using a specified statistical test at a given significance level (e.g., $\alpha = 0.05$).

If T is the test statistic to be calculated and C is the critical region, then the (frequentist) sample size is chosen so that

$$P(T \in C \mid \delta = 0) = \alpha,$$

$$P(T \in C \mid \delta = \delta_a) = 1 - \beta.$$

Assuming $T \sim N(\delta, \sigma^2/n)$, where n is the number of patients per treatment group, we are led to

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta_a^2}. \quad (8.2)$$

In survival analysis with proportional hazards, the test statistic T is usually taken to be the log-rank statistic. The log-rank statistic is approximately normally distributed.

For group sequential clinical trials, n is adjusted by a suitable multiple depending on the type of boundary and the number of interim looks..

A similar formula can be obtained if we have binomial responses.

Bayesian Sample Size

In a purely Bayesian approach, a particular sample size need not be chosen. The only decision required is whether to start the trial. The decision depends on the costs and benefits of the two choices. Once the trial has begun, decisions are made (at any time) whether to continue. However, such an approach may not be feasible for some trials.

There has been some work done for Bayesian sample size determination. Articles include Adcock (1987, 1992, 1995), Pham-Gia and Turkkan (1992), Pham-Gia (1995), Joseph et al. (1995a, 1995b), Joseph and Belisle (1997), and Spiegelhalter and Freedman (1986). Most of these papers can be found in *The Statistician*, which is also known as JRSS-D. The 1997 issue # 2 of *The Statistician* is devoted to sample size calculations.

Three common methods are the i) Average Coverage Criterion (ACC), ii) Average Length Criterion (ALC), and the iii) Worst Outcomes Criterion (WOC).

Average Coverage Criterion (ACC)

For a fixed posterior interval of length l , we can determine the sample size by finding the smallest n such that the equation

$$\int \left(\int_a^{a+l} p(\delta|x) d\delta \right) p(x) dx \geq 1 - \alpha$$

is satisfied. Here,

$$p(x) = \int p(x|\delta)\pi(\delta) d\delta$$

is the prior predictive distribution of x . This Average Coverage Criterion (ACC) ensures that the mean coverage of posterior credible intervals of length l , weighted by $p(x)$, is at least $1 - \alpha$.

Adcock (1988) first proposed the use of ACC in the context of estimating normal means, where the interval $(a, a + l)$ was chosen to be a symmetric tolerance region around the mean. Joseph et al. (1995) proposed that the interval $(a, a + l)$ be chosen to be a Highest Posterior Density (HPD) interval for asymmetric posterior distributions.

Average Length Criterion (ALC)

For a fixed posterior credible interval of coverage $1 - \alpha$, we can also determine the sample size by finding the smallest n such that

$$\int l^*(x, n) p(x) dx \leq l$$

where $l^*(x, n)$ is the length of the $100(1 - \alpha)\%$ posterior credible interval for data x , determined by solving

$$\int_a^{a+l^*(x, n)} p(\delta|x) d\delta = 1 - \alpha$$

for $l^*(x, n)$ for each value of $x \in \mathcal{X}$. As before, a can be chosen to give HPD intervals or symmetric intervals.

The ALC ensures that the mean length of the $100(1 - \alpha)\%$ posterior credible intervals weighted by $p(x)$ is at most l . Since most researchers will report intervals of fixed coverage (usually 95%) regardless of their length, it can be argued that the ALC is more conventional than ACC.

Worst Outcome Criterion (WOC)

Cautious investigators may not be satisfied with the “average” assurances provided by the ACC and the ALC criteria. Therefore, a conservative sample size can also be determined by

$$\inf_{x \in \mathcal{X}^*} \left(\int_a^{a+l(x,n)} p(\delta|x) d\delta \right) \geq 1 - \alpha,$$

where \mathcal{X}^* is a suitably chosen subset of the sample space \mathcal{X} .

For example, the WOC ensures that if \mathcal{X}^* consists of the most likely 95% of the possible $x \in \mathcal{X}$, then there is a 95% assurance that the length of the 100($1 - \alpha$)% posterior credible interval will be at most l .

Example 8.1: Sample Sizes for a Single Normal Mean

Let us first consider a one-sample problem, in which we have **one** normal population. Suppose x_1, \dots, x_n are i.i.d. $N(\mu, \sigma^2)$. Let $\tau = 1/\sigma^2$.

Assume the usual conjugate priors for (μ, τ) . That is,

$$\mu | \tau \sim N\left(\mu_0, \frac{1}{n_0 \tau}\right)$$

and

$$\tau \sim \text{gamma}(\alpha_0, \lambda_0).$$

Case 1: τ known

In the case that τ is known (i.e., no prior needed for τ), we have

$$\mu|x \sim N(\mu_n, \tau_n^{-1}),$$

where

$$\mu_n = \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}$$

and

$$\tau_n = (n + n_0)\tau.$$

Since the posterior precision depends only on n and does not vary with the observed data vector x , all three criteria (ACC, ALC, WOC) lead to the same solution, which is

$$n \geq \frac{4z_{1-\alpha/2}^2}{\tau l^2} - n_0, \quad (8.3)$$

where l = length of the desired posterior interval for μ . If a uniform improper prior for μ is used (i.e., $n_0 = 0$), then inequality (8.3) reduces to the frequentist formula for the sample size.

Case 2: τ unknown

If τ is unknown, then

$$\mu|x \sim t \left(n + 2\alpha_0, \mu_n, \frac{2\beta_n}{(n + 2\alpha_0)(n + n_0)} \right),$$

where

$$\beta_n = \lambda_0 + \frac{n}{2}s^2 + \frac{nn_0}{2(n + n_0)}(\bar{x} - \mu_0)^2$$

and $ns^2 = \sum_{i=1}^n (x_i - \bar{x})^2$.

Since the posterior precision varies with the data x , different criteria will lead to different sample sizes.

For the Average Coverage Criterion (ACC), the formula for the sample size is given by

$$n = \left(\frac{4\lambda_0}{\alpha_0 l^2} \right) t_{(2\alpha_0, 1-\alpha/2)}^2 - n_0. \quad (8.4)$$

Since α_0/λ_0 is the prior mean for τ , the ACC sample size for unknown τ is similar to that for known τ , in that we only need to substitute the prior mean for τ in inequality (8.3) and exchange the normal quantile z with a quantile from a $t_{2\alpha_0}$ distribution. Since the degrees of freedom of the t distribution do not increase with the sample size, equation (8.4) can lead to sample sizes that are substantially different from those in (8.2) or (8.3).

For the Average Length Criterion (ALC), it can be shown that the required sample size satisfies

$$2t_{(n+2\alpha_0, 1-\alpha/2)} \left(\frac{2\lambda_0}{(n + 2\alpha_0)(n + n_0)} \right)^{1/2} \frac{\Gamma(\frac{n+2\alpha_0}{2}) \Gamma(\frac{2\alpha_0-1}{2})}{\Gamma(\frac{n+2\alpha_0-1}{2}) \Gamma(\alpha_0)} \leq l. \quad (8.5)$$

For more details on this, see Joseph and Belisle (1997, JRSS-D). Although it does not appear feasible to solve the inequality in (8.5) explicitly for n , it is straightforward to calculate given α_0 , λ_0 , n_0 , α , and l .

For the Worst Outcomes Criterion (WOC), it can be shown that n satisfies

$$\frac{l^2(n + 2\alpha_0)(n + n_0)}{8\lambda_0 \left(1 + \frac{n}{2\alpha_0} F_{(n, 2\alpha_0, 1-\alpha)}\right)} \geq t_{(n+2\alpha_0, 1-\alpha/2)}^2, \quad (8.6)$$

where $F(n, 2\alpha_0, 1 - \alpha)$ denotes the $100(1 - \alpha)$ percentile of the F distribution with $(n, 2\alpha_0)$ degrees of freedom.

The smallest n satisfying (8.5) or (8.6) can be found by a search algorithm. See Joseph and Belisle (1997, JRSS-D) for more details.

Table 8.1: Sample sizes based on the various criteria

Example	α_0	λ_0	n_0	l	$1 - \alpha$	freq*	ACC	ALC	WOC
1	2	2	10	0.5	0.99	107	330	160	589
2	2	2	10	0.2	0.95	385	761	595	2152
3	2	2	10	0.2	0.80	165	226	248	914
4	2	2	10	0.2	0.50	46	45	61	245
5	100	100	100	0.2	0.95	385	289	288	344
6	100	100	10	0.2	0.95	385	379	378	436

* The frequentist sample size satisfies

$$n \geq \frac{4z_{1-\alpha/2}^2}{\tau l^2},$$

(i.e., equation (8.3) with $n_0 = 0$).

Examples 1-3 in Table 8.1 show that the Bayesian approach can provide larger sample sizes than the frequentist approach, even though the prior information is incorporated in the final inferences.

The same examples also illustrate that the sample size provided by the ALC tends to be smaller than that of the ACC when $1 - \alpha$ is near 1 and l is not near 0. This is because coverage probabilities are bounded above by 1, so that maintaining the required average coverage becomes more difficult as $1 - \alpha$ becomes larger.

Similarly, since l is bounded below by 0, maintaining an average length of l becomes more difficult as l approaches 0, leading to the large sizes for the ALC compared with the ACC in Example 4 in the table.

Example 5 shows that with a large amount of prior information on both (μ, τ) , the Bayesian approach leads to smaller sample sizes than the frequentist approach.

With an informative prior on τ , but not on μ , similar sample sizes are provided by all criteria, as suggested by Example 6, with the WOC criterion somewhat higher than the rest.

Additional Bayesian criteria for determining sample size include the following:

- 1) Determine n such that the average (or maximum) posterior variance of δ is equal to some prespecified quantity. The averaging (or maximizing) is done with respect to $p(x)$. For example, let

$$\psi(n) = \int_{\mathcal{X}} \text{Var}(\delta | x) p(x) dx.$$

Let ϵ be the prescribed value for $\psi(n)$. Then we set $\epsilon = \psi(n)$ and solve for n .

- 2) Choose the sample size so that the Bayes risk is less than or equal to some prespecified quantity. This procedure requires specifying a loss function.

- 3) Choose the sample size for a chosen value of the predictive power.

$$p^*(n) = \int_{\delta_i}^{\infty} P(\text{reject } H_0 \mid \delta) p(\delta) d\delta.$$

When the prior is a point mass, then $p^*(n)$ reduces to the frequentist power.

- 4) There is some recent work on determining sample size from Bayes factors.

Monitoring a Trial

Most Phase III trials consist of group sequential trials, in which several analyses of the data, called interim analyses, are conducted at various stages of the trial to determine the effects of the treatments involved.

For the Bayesian, the relevant quantity for monitoring the trial is the posterior distribution of δ , $p(\delta | x)$. HPD intervals are constructed for δ to examine the behavior of the treatment given the current data.

Falsehood #1: “From the Bayesian perspective, a predetermined number of looks is not relevant since the significance level is not relevant. The statistician can look at the data as many times as they wish.”

Falsehood #2: “An important aspect of the Bayesian approach is that all available information can be used in deciding whether to stop the trial. There is no penalty for continual data analysis.”

Example 8.2

We consider an example of a group sequential design for a two-arm Phase III cancer clinical trial where the primary endpoint is overall survival.

The median overall survival for the standard treatment arm is 8 months, and the overall survival for the new treatment arm is hypothesized to be 12 months. We want to find the sample size to detect this difference with at least 80% power using a two-sided log-rank test with overall significance level of 0.05.

Assume 25 months of accrual and 17 months of followup with an accrual rate of 12 patients per month. This leads to 300 patients which yields 89% power.

Real Time (months)	Information Time	Deaths Under H_1	Upper Bound	Rejection Probability Under H_0	Nominal Significance Level	Rejection Probability Under H_1
14	0.26	67	4.2432	< 0.001	< 0.001	0.005
21	0.50	128	2.9681	0.003	0.003	0.244
30	0.80	208	2.2494	0.023	0.024	0.500
42	1.00	256	2.0337	0.024	0.042	0.143
Total				0.05		0.892

Prior for δ

The choice of prior distributions is an important issue in Bayesian inference.

If δ denotes a mean treatment difference, log hazard ratio, or a difference in two proportions, a normal prior can be taken for it. If δ is a proportion, a beta prior can be chosen. In the normal case, we have

$$p(\delta) = N(\delta_0, \sigma^2/n_0),$$

where δ_0 is the prior mean. This prior is equivalent to a normalized likelihood arising from a hypothetical trial of n_0 patients with an observed value δ_0 of the treatment difference statistic.

When reporting studies, we should acknowledge that different individuals or groups hold different prior beliefs. Thus, we may consider different types of priors, such as reference, clinical, skeptical, and enthusiastic priors.

Reference priors: Reference priors are supposed to represent minimal information, and they can be obtained by letting $n_0 \rightarrow 0$.

Clinical priors: A clinical prior is intended to formalize opinion of well-informed specific individuals, often those taking part in the trial themselves. Deriving such a prior (e.g., (δ_0, n_0)) requires asking lots of questions and/or having data from previous studies.

Skeptical priors: Formalization of the prior belief that large differences are unlikely. Such an opinion could be represented by taking $\delta_0 = 0$ with a suitable n_0 . Thus we express skepticism concerning δ by choosing a prior distribution which is normal with mean 0 and such that $P(\delta > \delta_a)$ is a small value, say γ .

Enthusiastic priors: This prior represents individuals who are reluctant to stop when results supporting the null hypothesis are observed. Such a prior may have mean δ_a , and $P(\delta < 0) = \gamma$.

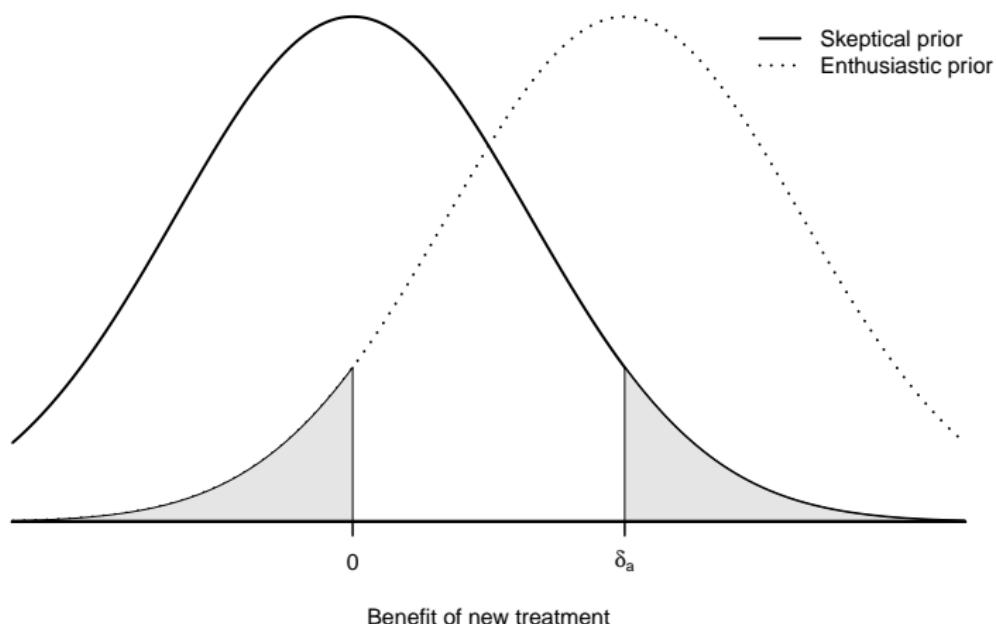


Figure 8.2: Skeptical and enthusiastic priors for a trial with alternative hypothesis δ_a . The skeptics' probability that the true difference is greater than δ_a is γ (right shaded tail), which is also the enthusiasts' probability that the true difference is less than 0.

Evidence for clinical priors can come from any of the following sources:

- 1) Evidence from other randomized trials - if similar previous studies are available, then that information could be used as a basis for a prior distribution.
- 2) Evidence from non-randomized studies - this information could possibly be used, but one may want to specify a large prior variance for δ in this case.
- 3) Subjective clinical opinion - this includes interviews with clinicians involved in the trial, and it is the most subjective prior elicitation.

Predictions

A major strength of the Bayesian approach is the ease of making predictions concerning the events of interest. Thus, a Bayesian approach allows for calculating the probability that a future patient will respond to therapy.

Probabilities of future observations are much more difficult in a formal classical (frequentist) approach.

Suppose m observations have been collected and we are interested in the possible consequences of continuing the trial for a further n observations. If we denote our future statistic by t_n , then the predictive distribution of t_n has the form

$$p_m(t_n) = \int_{\Delta} p(t_n | \delta) p(\delta | x_m) d\delta$$

where $p(\delta | x_m)$ is the posterior distribution of δ based on m observations and the statistic x_m .

Example 8.3: Levamisole and 5-fluorouracil in Bowel Cancer

Moertel et al. (1990, NEJM) have reported results from a randomized Phase III cancer clinical trial investigating the effect of the drug levamisole (LEV) alone or in combination with 5-fluorouracil (5-FU) for patients with resected cancer of the colon or rectum.

Patients entering the trial were allocated to one of three treatment arms: LEV, LEV + 5-FU, or control.

The study was designed with an alternative $\delta_a = \log(1.35) = .30$, and to have 90% power for a one-sided test with 5% significance, and thus required 380 deaths. The log hazard was based on the comparison between LEV + 5-FU against control.

On a log-hazard-ratio scale, the range of equivalence is $(\delta_i, \delta_s) = (0, .29)$.

Priors: We specify both a skeptical prior and an enthusiastic prior, as shown in Figure 8.3 (a). For the skeptical prior, we set $\delta_0 = 0$ and $P(\delta > .30) = .05$, and setting $\sigma = 2$ gives $n_0 = 120$. The enthusiastic prior has mean $.30$ and the same variance.

Likelihood: In an interim analysis for LEV + 5-FU vs. control, the observed log-hazard ratio was $x_m = .40$ with $m = 192$ total deaths. The reference posterior (normalized likelihood) in Figure 8.3 (b) shows convincing evidence that $\delta > 0$ (probability = $.003$), but moderate evidence that the treatment is clinically superior (probability = $.777$). The predictive distribution shows that there was only a 4% chance of observing such a high result given the skeptical prior (Figure 8.3 (d)).

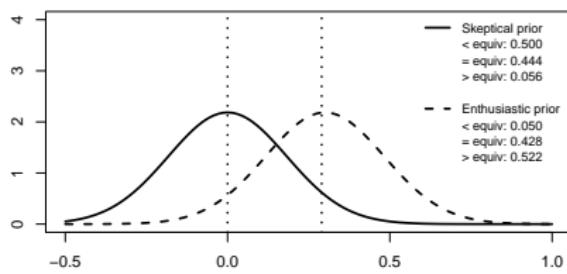
Posterior distributions: The skeptical posterior distribution in Figure 8.3 (c) has mean $.25$ and standard deviation $.11$, corresponding to a 95% HPD interval of $(1.03, 1.60)$ for the hazard ratio.

There is a small posterior probability (.015) of the treatment difference being less than 0 (i.e., of the control treatment being superior to LEV + 5-FU), whereas the posterior mean is within the range of equivalence.

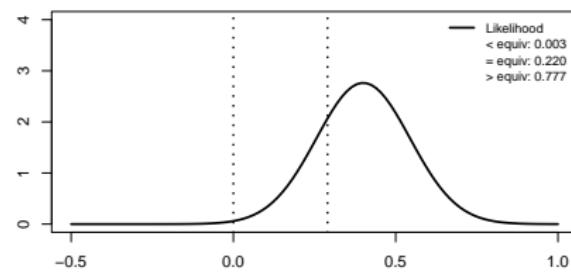
Results are available from a previous study which showed an apparent reduction in death-rate of patients receiving LEV + 5-FU, with an estimated log-hazard ratio of .14 with standard error .17.

With these data included into the analysis and using a skeptical prior, the hazard ratio estimate is changed to 1.25 with 95% interval (1.04, 1.49).

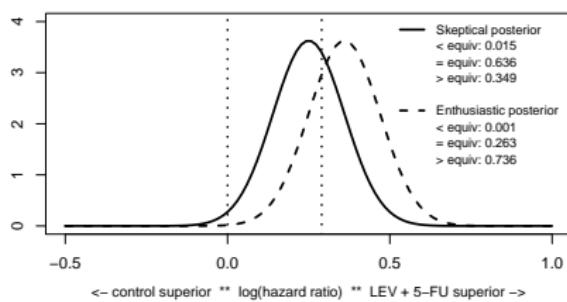
Including the previous data into the analysis increases certainty that $\delta > 0$ (probability = .007 with the skeptical prior) but decreases the certainty that the treatment is clinically superior, $p(\delta > .29 \mid \text{combined evidence}) = .22$.



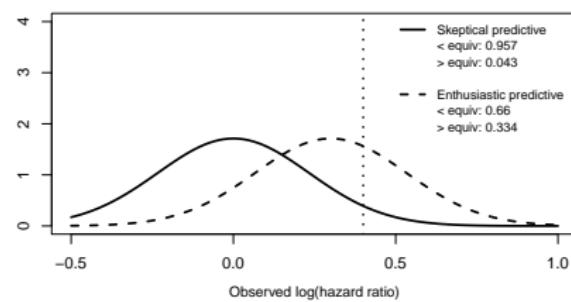
(a)



(b)



(c)



(d)

Figure 8.3: Prior, likelihood and posterior distributions for LEV + 5-FU versus control (the probabilities of falling below, within and above the range of equivalence are shown in the top right-hand corner, first for the skeptical prior and then for the enthusiastic prior): (a) prior; (b) likelihood ($m = 192$; $x_m = 0.4$); (c) posterior; (d) predictive distribution.

The investigators' original plan was to include approximately 188 further deaths in the comparison of control and LEV + 5-FU. Table 8.2 shows the predictive probability of different conclusions being drawn at the end of this period, calculated under three assumptions:

- i) a fully Bayesian analysis using our skeptical prior,
- ii) that the prior is used for predictions but excluded from the final analysis, and
- iii) that the predictions and analysis are based solely on the data (equivalent to using the reference prior throughout).

Table 8.2: Predictive probability of the final 99% interval for the hazard of control relative to LEV + 5-FU having different positions relative to the range of equivalence, after observation of a further 188 events

Position of 99% interval	Results for the following modes of prediction:		
	Bayesian	Mixed	Likelihood
A: old superior	0.000	0.000	0.000
B: new not superior	0.004	0.001	0.000
C: equivocal	0.407	0.256	0.091
D: old not superior	0.590	0.737	0.845
E: new superior	0.000	0.006	0.064

See Figure 8.1 for illustration of A, B, C, D, and E.

We see that ignoring the prior completely leads to a final interval that is very likely to exclude $\delta = 0$ (D or E), which might reinforce the decision to stop the trial.

If the prior is used for prediction, then the chance that a final 99% interval includes 0 rises to .26, whereas the chance that the skeptical posterior interval includes 0 is .41.

Thus the eventual result of this study is not a foregone conclusion.

Advantages of Bayesian Approach

The Bayesian approach has several advantages:

1) Direct probabilities

Bayesian probabilities are direct and frequentist probabilities are not.

2) Using all the evidence

Bayesian inferences are conditioned on all available information while frequentist measures are specific to a particular experiment and to that experiment's design. Bayesian inferences depend only on data actually observed while frequentist measures involve probabilities of data that were possible given the design of the trial but were not observed (i.e., p-values).

3) Flexibility

Bayesian inferences can be updated continually as data accumulate (i.e., sequential monitoring).

4) Predictive probabilities

A Bayesian approach allows for calculating predictive probabilities. Probabilities of future observations are much more difficult in a formal frequentist approach.

5) Decision making

The Bayesian approach is tailored to decision making. Designing a clinical trial is a decision problem. Allocating resources among various research projects is a decision problem. Stopping drug development is a decision problem. There are costs and benefits in these problems. In the Bayesian approach, these costs and benefits and the available information are assessed. The costs and benefits are then weighed by the predictive probabilities from each possible decision.

Chapters of *Bayesian Survival Analysis*

The following chapters come from *Bayesian Survival Analysis* (Ibrahim, Chen, and Sinha; Springer):

Chapter 1: Introduction

Survival analysis, the Bayesian paradigm.

Chapter 2: Parametric Models

Exponential model, Weibull model, extreme value model, log-normal model, gamma model.

Chapter 3: Semiparametric Models

Piecewise constant hazard model, models using a gamma process, beta process models, Dirichlet process models.

Chapter 4: Frailty Models

$$\begin{aligned} h(y) &= h_0(y) \exp\{x_i' \beta + z_i' b_i\} \\ b_i &\sim N(0, \sigma^2) \\ h(y_i) &= h_0(y_i) w_i \exp\{x_i' \beta\} \\ w_i &\sim \text{gamma}(\eta, \eta) \end{aligned}$$

Chapter 5: Cure Rate Models

proper survival model: $\lim_{t \rightarrow \infty} S(t) = 0$ cure rate model: $\lim_{t \rightarrow \infty} S(t) = p$, where $p > 0$

Chapter 6: Model Comparison

BIC, CPO, L measure, Bayes factors, posterior model probabilities for survival data.

Chapter 7: Joint Models for Longitudinal and Survival Data

We have a longitudinal marker such as CD4 count, immune response to a vaccine, or quality of life (QOL). We wish to examine the association between the longitudinal marker and a time to event. The model for the longitudinal marker is a random effects model, which is then “linked” to the survival model.

Chapter 8: Missing Covariate Data

Bayesian methods for handling missing covariate data in survival models.

Chapter 9: Design and Monitoring of Randomized Clinical Trials

Chapter 10: Other Topics

Non-proportional hazards, AFT models, residuals, diagnostics.