
Half Title

Solution Manual
Introduction to Bayesian Inference:
A GUIDed tour using R



Title Page

Solution Manual
Introduction to Bayesian Inference:
A GUIDed tour using R

by Andrés Ramírez-Hassan, PhD. Statistical Science.



To my parents, Nancy and Orlando.



Contents

Foreword	ix
Preface	xi
Symbols	xiii
I Foundations: Theory, simulation methods and programming	1
1 Basic formal concepts	3
1.1 Solutions of Exercises	3
2 Conceptual differences of the Bayesian and Frequentist approaches	23
2.1 Solutions of Exercises	23
3 Objective and subjective Bayesian approaches	29
3.1 Solutions of Exercises	29
4 Cornerstone models: Conjugate families	31
4.1 Solutions of Exercises	31
5 Simulation methods	55
5.1 Solutions of Exercises	55
6 Univariate models	57
6.1 Solutions of Exercises	57
Bibliography	65





Foreword





Preface



Symbols

Symbol Description

\neg	Negation symbol.	\mathcal{R}	The Real set.
\propto	Proportional symbol.	\emptyset	Empty set.
\perp	Independence symbol.	$\mathbb{1}$	Indicator function.



Part I

Foundations: Theory, simulation methods and programming



1

Basic formal concepts

1.1 Solutions of Exercises

1. *The court case: the blue or green cap*

A cab was involved in a hit and run accident at night. There are two cab companies in the town: blue and green. The former has 150 cabs, and the latter 850 cabs. A witness said that a blue cab was involved in the accident; the court tested his/her reliability under the same circumstances, and got that 80% of the times the witness correctly identified the color of the cab. *What is the probability that the color of the cab involved in the accident was blue given that the witness said it was blue?*

Answer

Set WB and WG equal to the events that the witness said the cab was blue and green, respectively. Set B and G equal to the events that the cabs are blue and green, respectively. We need to calculate $P(B|WB)$, then:

$$\begin{aligned} P(B|WB) &= \frac{P(B, WB)}{P(WB)} \\ &= \frac{P(WB|B) \times P(B)}{P(WB|B) \times P(B) + (1 - P(WB|B)) \times (1 - P(B))} \\ &= \frac{0.8 \times 0.15}{0.8 \times 0.15 + 0.2 \times 0.85} \\ &= 0.41 \end{aligned} \tag{1.1}$$

2. *The Monty Hall problem*

What is the probability of winning a car in the *Monty Hall problem* switching the decision if there are four doors, where there are three goats and one car? Solve this problem analytically and computationally. What if there are n doors, $n - 1$ goats and one car?

Answer

Let's name P_i the event *contestant picks door No. i* , H_i the event *host picks door No. i* , and C_i the event *car is behind door No. i* . Let's assume that the

contestant picked door number 1, and the host picked door number 3, then the contestant is interested in the probability of the event $P(C_i|H_3, P_1)$, $i = 2$ or 4 . Then, $P(H_3|C_3, P_1) = 0$, $P(H_3|C_2, P_1) = P(H_3|C_4, P_1) = 1/2$ and $P(H_3|C_1, P_1) = 1/3$. Then,

$$\begin{aligned}
 P(C_i|H_3, P_1) &= \frac{P(C_i, H_3, P_1)}{P(H_3, P_1)} \\
 &= \frac{P(H_3|C_i, P_1)P(C_i|P_1)P(P_1)}{P(H_3|P_1) \times P(P_1)} \\
 &= \frac{P(H_3|C_i, P_1)P(C_i)}{P(H_3|P_1)} \\
 &= \frac{1/2 \times 1/4}{1/3} \\
 &= \frac{3}{8},
 \end{aligned} \tag{1.2}$$

where the third equation uses the fact that C_i and P_i are independent events, and $P(H_3|P_1) = 1/3$ due to this depending just on P_1 (not on C_i).

Therefore, changing the initial decision increases the probability of getting the car from $1/4$ to $3/8$!

Let's check the case with n doors, and assume that the contestant picks the door No. 1, the car is behind the door No. n , and the host, who knows what is behind each door, opens any of the remaining $n - 2$ doors, where there is a goat. The contestant is interested in the probability of the event:

$$\begin{aligned}
 P(C_n|(H_2 \cup \dots \cup H_{n-1}) \cap P_1) &= \frac{P((H_2 \cup H_3 \cup \dots \cup H_{n-1})|C_n \cap P_1)P(C_n|P_1)P(P_1)}{P((H_2 \cup H_3 \cup \dots \cup H_{n-1})|P_1)P(P_1)} \\
 &= \frac{[P(H_2|C_n \cap P_1) + \dots + P(H_{n-1}|C_n \cap P_1)]P(C_n)}{P(H_2|P_1) + P(H_3|P_1) + \dots + P(H_{n-1}|P_1)} \\
 &= \frac{1 \times (\frac{1}{n})}{\frac{1}{n-1} + \frac{1}{n-1} + \dots + \frac{1}{n-1}} \\
 &= \left(\frac{1}{n}\right) \left(\frac{n-1}{n-2}\right).
 \end{aligned} \tag{1.3}$$

In general, the probability of winning the car changing the pick is $\frac{1}{n} \frac{n-1}{n-2}$, while the probability of winning given no change is $\frac{1}{n}$. Given that $\frac{1}{n} \frac{n-1}{n-2} > \frac{1}{n}$ for all $n \geq 3$, where the difference between both probabilities is $\frac{1}{n(n-2)}$. We observe that as the number of doors increases, the difference between the two probabilities becomes zero.

Let's see a code for the general setting,

R code. The Monty Hall Problem

```

1 set.seed(0101) # Set simulation seed
2 S <- 100000 # Simulations
3 Game <- function(opt = 3){
4   # opt: number of options. opt > 2
5   opts <- 1:opt
6   car <- sample(opts, 1) # car location
7   guess1 <- sample(opts, 1) # Initial guess
8   if(opt == 3 && car != guess1) {
9     host <- opts[-c(car, guess1)]
10    } else {
11    host <- sample(opts[-c(car, guess1)], 1)
12  }
13  win1 <- guess1 == car # Win given no change
14  if(opt == 3) {
15    guess2 <- opts[-c(host, guess1)]
16  } else {
17    guess2 <- sample(opts[-c(host, guess1)], 1)
18  }
19  win2 <- guess2 == car # Win given change
20  return(c(win1, win2))
21 }
22 #Win probabilities
23 Prob <- rowMeans(replicate(S, Game(opt = 4)))
24 #Winning probabilities no changing door
25 Prob[1]
26 0.25151
27 #Winning probabilities changing door
28 Prob[2]
29 0.37267

```

3. Solve the health insurance example using a Gamma prior in the rate parametrization, that is, $\pi(\lambda) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} \exp\{-\lambda\beta_0\}$.

Answer

First, we get the posterior distribution,

$$\pi(\lambda|\mathbf{y}) = \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} e^{-\lambda\beta_0} \right) \left(\prod_{i=1}^N \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \right) \quad (1.4)$$

$$\begin{aligned}
&= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} e^{-\lambda\beta_0} \right) \left(\frac{\lambda^{\sum_{i=1}^N y_i} e^{-N\lambda}}{\prod_{i=1}^N y_i!} \right) \\
&= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{1}{\prod_{i=1}^N y_i!} \right) \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} e^{-\lambda(\beta_0 + N)} \\
&\propto \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} e^{-\lambda(\beta_0 + N)}.
\end{aligned} \tag{1.5}$$

The last expression is the kernel of a Gamma distribution with parameters $\alpha_n = \sum_{i=1}^N y_i + \alpha_0$ and $\beta_n = \beta_0 + N$.

Given that $\int_0^\infty \pi(\lambda|\mathbf{y}) d\lambda = 1$, then the constant of proportionality in the last expression is $\Gamma(\alpha_n) / \beta_n^{\alpha_n}$. Therefore the posterior density function $\pi(\lambda|\mathbf{y})$ is $G(\alpha_n, \beta_n)$.

The posterior mean is

$$\begin{aligned}
E[\lambda|\mathbf{y}] &= \frac{\alpha_n}{\beta_n} \\
&= \frac{\sum_{i=1}^N y_i + \alpha_0}{\beta_0 + N} \\
&= \left(\frac{N}{\beta_0 + N} \right) \bar{y} + \left(\frac{\beta_0}{\beta_0 + N} \right) \frac{\alpha_0}{\beta_0} \\
&= w\bar{y} + (1-w) E[\lambda],
\end{aligned} \tag{1.6}$$

where $w = \frac{N}{\beta_0 + N}$, \bar{y} is the sample mean, and $E[\lambda] = \frac{\alpha_0}{\beta_0}$.

The posterior predictive distribution is given by

$$\begin{aligned}
\pi(Y_0|\mathbf{y}) &= \int_0^\infty \frac{\lambda^{y_0} e^{-\lambda}}{y_0!} \pi(\lambda|\mathbf{y}) d\lambda \\
&= \int_0^\infty \left(\frac{\lambda^{y_0} e^{-\lambda}}{y_0!} \right) \left(\frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \lambda^{\alpha_n-1} e^{-\lambda\beta_n} \right) d\lambda \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n) y_0!} \int_0^\infty \lambda^{y_0 + \alpha_n - 1} e^{-\lambda(1+\beta_n)} d\lambda \\
&= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n) y_0!} \frac{\Gamma(y_0 + \alpha_n)}{(1 + \beta_n)^{y_0 + \alpha_n}} \\
&= \frac{\Gamma(y_0 + \alpha_n)}{\Gamma(\alpha_n) y_0!} \left(\frac{1}{1 + \beta_n} \right)^{y_0} \left(\frac{\beta_n}{1 + \beta_n} \right)^{\alpha_n}
\end{aligned} \tag{1.7}$$

$$\begin{aligned}
&= \frac{(y_0 + \alpha_n - 1)!}{(\alpha_n - 1)! y_0!} \left(\frac{1}{1 + \beta_n} \right)^{y_0} \left(\frac{\beta_n}{1 + \beta_n} \right)^{\alpha_n} \\
&= \binom{y_0 + \alpha_n - 1}{y_0} \left(\frac{1}{1 + \beta_n} \right)^{y_0} \left(\frac{\beta_n}{1 + \beta_n} \right)^{\alpha_n}.
\end{aligned}$$

Therefore $Y_0|y \sim NB(\alpha_n, p_n)$ where $p_n = \frac{1}{1+\beta_n}$.

To use empirical Bayes, we have the following setting

$$[\hat{\alpha}_0, \hat{\beta}_0] = \arg \max_{\alpha_0, \beta_0} \ln(p(\mathbf{y})),$$

where

$$\begin{aligned}
p(y) &= \int_0^\infty \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} e^{-\lambda\beta_0} \right) \left(\prod_{i=1}^N \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \right) d\lambda \quad (1.8) \\
&= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0) \prod_{i=1}^N y_i!} \right) \int_0^\infty \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} e^{-\lambda(\beta_0 + N)} d\lambda \\
&= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0) \prod_{i=1}^N y_i!} \right) \left(\frac{\Gamma(\sum_{i=1}^N y_i + \alpha_0)}{(\beta_0 + N)^{\sum_{i=1}^N y_i + \alpha_0}} \right) \\
&= \frac{\Gamma(\sum_{i=1}^N y_i + \alpha_0)}{\Gamma(\alpha_0) \prod_{i=1}^N y_i!} \left(\frac{1}{\beta_0 + N} \right)^{\sum_{i=1}^N y_i} \left(\frac{\beta_0}{\beta_0 + N} \right)^{\alpha_0}.
\end{aligned}$$

*R code. Health insurance, predictive distribution
using vague hyperparameters*

```

1 set.seed(010101)
2 y <- c(0, 3, 2, 1, 0) # Data
3 N <- length(y)
4
5 # Predictive distribution
6 ProbBo <- function(y, a0, b0){
7   N <- length(y)
8   #sample size
9   aN <- a0 + sum(y)
10  # Posterior shape parameter
11  bN <- b0 + N
12  # Posterior scale parameter
13  p <- 1 / (bN + 1)
14  # Probability negative binomial density
15  Pr <- 1 - pbinom(0, size = aN, prob = (1 - p))
16  # Probability of visiting the Doctor
17  # Observe that in R there is a slightly
18  # different parametrization.
19  return(Pr)
20 }
21
22 # Using a vague prior:
23 a0 <- 0.001 # Prior shape parameter
24 b0 <- 0.001 # Prior scale parameter
25 PriMeanV <- a0 / b0 # Prior mean
26 PriVarV <- a0 / b0^2 # Prior variance
27 Pp <- ProbBo(y, a0 = 0.001, b0 = 0.001)
28 # This setting is vague prior information.
29 Pp
30 0.67

```

*R code. Health insurance, predictive distribution
using empirical Bayes*

```

1 # Using Empirical Bayes
2 LogMgLik <- function(theta, y){
3   N <- length(y)
4   #sample size
5   a0 <- theta[1]
6   # prior shape hyperparameter
7   b0 <- theta[2]
8   # prior scale hyperparameter
9   aN <- sum(y) + a0
10  # posterior shape parameter
11  if(a0 <= 0 || b0 <= 0){
12    #Avoiding negative values
13    lnp <- -Inf
14  }else{lnp <- lgamma(aN) - sum(y)*log(b0+1) + a0*log(b0/(b0
15    +1)) - lgamma(a0)}
16    # log marginal likelihood
17    return(-lnp)
18  }
19  theta0 <- c(0.01, 0.01)
20  # Initial values
21  control <- list(maxit = 1000)
22  # Number of iterations in optimization
23  EmpBay <- optim(theta0, LogMgLik, method = "BFGS", control =
24    control, hessian = TRUE, y = y)
25  # Optimization
26  EmpBay$convergence
27  # Checking convergence
28  EmpBay$value # Maximum
29  a0EB <- EmpBay$par[1]
30  # Prior shape using empirical Bayes
31  a0EB
32  128.383
33  b0EB <- EmpBay$par[2]
34  # Prior scale using empirical Bayes
35  b0EB
36  106.801
37  PriMeanEB <- a0EB / b0EB
38  # Prior mean
39  PriVarEB <- a0EB / b0EB^2
40  # Prior variance
41  PpEB <- ProbBo(y, a0 = a0EB, b0 = b0EB)
42  # This setting is using empirical Bayes.
43  PpEB
44  0.69

```

4. Suppose that you are analyzing to buy a car insurance next year. To make

a better decision you want to know *what is the probability that you have a car claim next year?* You have the records of your car claims in the last 15 years, $\mathbf{y} = \{0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0\}$.

Assume that this is a random sample from a data generating process (statistical model) that is Bernoulli, $Y_i \sim \text{Ber}(p)$, and your probabilistic prior beliefs about p are well described by a beta distribution with parameters α_0 and β_0 , $p \sim B(\alpha_0, \beta_0)$, then, you are interested in calculating the probability of a claim the next year $P(Y_0 = 1|\mathbf{y})$.

Solve this using an empirical Bayes approach and a non-informative approach where $\alpha_0 = \beta_0 = 1$ (uniform distribution).

Answer

The posterior distribution is given by

$$\begin{aligned} \pi(p|\mathbf{y}) &= \left[\frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \right] \left[\prod_{i=1}^N p^{y_i} (1-p)^{1-y_i} \right] \quad (1.9) \\ &= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p^{\sum_{i=1}^N y_i + \alpha_0 - 1} (1-p)^{\beta_0 + N - \sum_{i=1}^N y_i - 1} \\ &\propto p^{\sum_{i=1}^N y_i + \alpha_0 - 1} (1-p)^{\beta_0 + N - \sum_{i=1}^N y_i - 1}. \end{aligned}$$

The last expression is the kernel of a Beta distribution with parameters $\alpha_n = \sum_{i=1}^N y_i + \alpha_0$ and $\beta_n = \beta_0 + N - \sum_{i=1}^N y_i$. Thus, the posterior mean is

$$\begin{aligned} E[p|\mathbf{y}] &= \frac{\alpha_n}{\alpha_n + \beta_n} \\ &= \frac{\sum_{i=1}^N y_i + \alpha_0}{\alpha_0 + \beta_0 + N} \quad (1.10) \\ &= \frac{N\bar{y}}{\alpha_0 + \beta_0 + N} + \frac{\alpha_0}{\alpha_0 + \beta_0 + N} \\ &= \frac{N}{\alpha_0 + \beta_0 + N} (\bar{y}) + \frac{\alpha_0 + \beta_0}{\alpha_0 + \beta_0 + N} \left(\frac{\alpha_0}{\alpha_0 + \beta_0} \right) \\ &= w(\bar{y}) + (1-w)E[p], \end{aligned}$$

where $w = \frac{N}{\alpha_0 + \beta_0 + N}$, \bar{y} is the sample mean, and $E[p] = \frac{\alpha_0}{\alpha_0 + \beta_0}$ is the prior mean.

The posterior predictive distribution of claim the next year is given by

$$\begin{aligned}
\pi(Y_0 = 1|\mathbf{y}) &= \int_0^1 P(Y_0 = 1|\mathbf{y}, p) \pi(p|\mathbf{y}) dp \\
&= \int_0^1 p \times \pi(p|\mathbf{y}) dp \\
&= \mathbb{E}[p|\mathbf{y}] \\
&= \frac{\alpha_n}{\alpha_n + \beta_n}.
\end{aligned} \tag{1.11}$$

To use empirical Bayes, we have the following setting

$$[\hat{\alpha}_0 \ \hat{\beta}_0] = \arg \max_{\alpha_0, \beta_0} \ln(p(\mathbf{y})),$$

where

$$\begin{aligned}
p(\mathbf{y}) &= \int_0^1 \left[\frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0) \Gamma(\beta_0)} p^{\alpha_0-1} (1-p)^{\beta_0-1} \right] \left[\prod_{i=1}^N (1-p)^{1-y_i} \right] dp \tag{1.12} \\
&= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0) \Gamma(\beta_0)} \int_0^1 p^{\sum_{i=1}^N y_i + \alpha_0 - 1} (1-p)^{\beta_0 + N - \sum_{i=1}^N y_i - 1} dp \\
&= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0) \Gamma(\beta_0)} \frac{\Gamma\left(\sum_{i=1}^N y_i + \alpha_0\right) \Gamma\left(\beta_0 + N - \sum_{i=1}^N y_i\right)}{\Gamma(\alpha_0 + \beta_0 + N)}.
\end{aligned}$$

R code. Car claim, predictive distribution using vague hyperparameters

```

1 set.seed(010101)
2 y <- c(0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0)
3 # Data
4 N <- length(y)
5 #require(TailRank)
6 # Predictive distribution
7 ProbBo <- function(y, a0, b0){
8   N <- length(y)
9   #sample size
10  aN <- a0 + sum(y)
11  # Posterior shape parameter
12  bN <- b0 + N - sum(y)
13  # Posterior scale parameter
14  pr <- aN / (aN + bN)
15  # Probability of a claim the next year
16  return(pr)
17 }
18 # Using a vague prior:
19 a0 <- 1 # Prior shape parameter
20 b0 <- 1 # Prior scale parameter
21 PriMeanV <- a0 / (a0 + b0)
22 # Prior mean
23 PriVarV <- (a0*b0) / (((a0+b0)^2)*(a0+b0+1))
24 # Prior variance
25 Pp <- ProbBo(y, a0 = 1, b0 = 1)
26 # This setting is defining vague prior information.
27 # The probability of a claim
28 Pp
29 0.47

```

R code. Car claim, predictive distribution using empirical Bayes

```

1 # Using Empirical Bayes
2 LogMgLik <- function(theta, y){
3   N <- length(y)
4   #sample size
5   a0 <- theta[1]
6   # prior shape hyperparameter
7   b0 <- theta[2]
8   # prior scale hyperparameter
9   aN <- sum(y) + a0
10  # posterior shape parameter
11  if(a0 <= 0 || b0 <= 0){
12    #Avoiding negative values
13    lnp <- -Inf
14  }else{lnp <- lgamma(a0+b0) + lgamma(aN) + lgamma(b0+N-sum(
15    y)) -lgamma(a0) - lgamma(b0) - lgamma(a0+b0+N)}
16  # log marginal likelihood
17  return(-lnp)
18 }
19 theta0 <- c(0.1, 0.1)
20 # Initial values
21 control <- list(maxit = 1000)
22 # Number of iterations in optimization
23 EmpBay <- optim(theta0, LogMgLik, method = "BFGS", control =
24   control, hessian = TRUE, y = y)
25 # Optimization
26 EmpBay$convergence
27 # Checking convergence
28 EmpBay$value # Maximum
29 a0EB <- EmpBay$par[1]
30 # Prior shape using empirical Bayes
31 b0EB <- EmpBay$par[2]
32 # Prior scale using empirical Bayes
33 PriMeanEB <- a0EB / (a0EB + b0EB)
34 # Prior mean
35 PriVarEB <- (a0EB*b0EB) / (((a0EB+b0EB)^2)*(a0EB+b0EB+1))
36 # Prior variance
37 PpEB <- ProbBo(y, a0 = a0EB, b0 = b0EB)
38 # This setting is using empirical Bayes.
39 PpEB
40 0.47

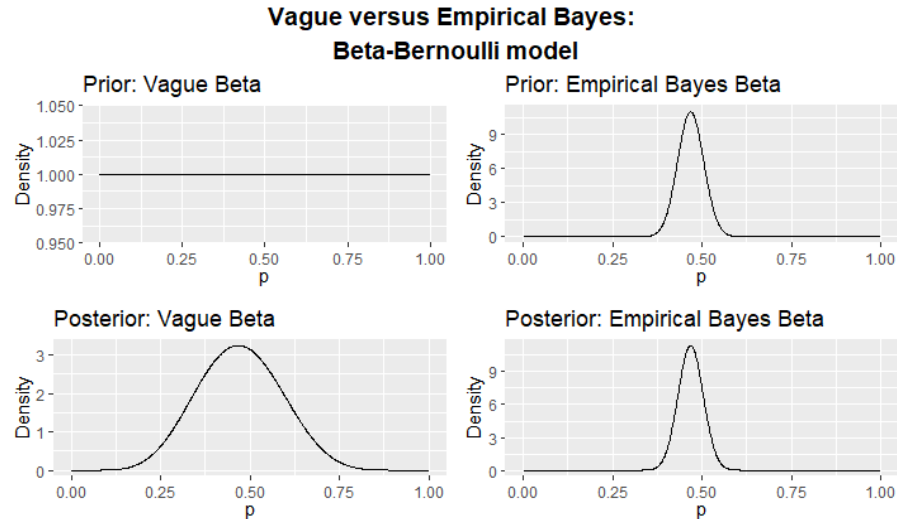
```

R code. Car claim, density plots

```

1 # Density figures
2 lambda <- seq(0.001, 1, 0.001)
3 # Values of lambda
4 VaguePrior <- dbeta(lambda, shape1 = a0, shape2 = b0)
5 EBPrior <- dbeta(lambda, shape1 = a0EB, shape2 = b0EB)
6 PosteriorV <- dbeta(lambda, shape1 = a0 + sum(y), shape2 =
  b0 + N - sum(y))
7 PosteriorEB <- dbeta(lambda, shape1 = a0EB + sum(y), shape2
  = b0EB + N - sum(y))
8 # Likelihood function
9 Likelihood <- function(theta, y){
10   LogL <- dbinom(y, 1, theta, log = TRUE)
11   # LogL <- dbern(y, theta)
12   Lik <- prod(exp(LogL))
13   return(Lik)
14 }
15 Liks <- sapply(lambda, function(par) {Likelihood(par, y = y)
  })
16 Sc <- max(PosteriorEB)/max(Liks)
17 #Scale for displaying in figure
18 LiksScale <- Liks * Sc
19 data <- data.frame(cbind(lambda, VaguePrior, EBPrior,
  PosteriorV, PosteriorEB, LiksScale))
20 #Data frame
21 require(ggplot2)
22 # Cool figures
23 require(latex2exp)
24 # LaTeX equations in figures
25 require(ggpubr)
26 # Multiple figures in one page
27 fig1 <- ggplot(data = data, aes(lambda, VaguePrior)) +
  geom_line() +
  xlab(TeX("$p$")) + ylab("Density") + ggtitle("Prior: Vague
  Beta")
28 fig2 <- ggplot(data = data, aes(lambda, EBPrior)) +
  geom_line() +
  xlab(TeX("$p$")) + ylab("Density") +
  ggtitle("Prior: Empirical Bayes Beta")
29 fig3 <- ggplot(data = data, aes(lambda, PosteriorV)) +
  geom_line() +
  xlab(TeX("$p$")) + ylab("Density") +
  ggtitle("Posterior: Vague Beta")
30 fig4 <- ggplot(data = data, aes(lambda, PosteriorEB)) +
  geom_line() +
  xlab(TeX("$p$")) + ylab("Density") +
  ggtitle("Posterior: Empirical Bayes Beta")
31 FIG <- ggarrange(fig1, fig2, fig3, fig4,
32 ncol = 2, nrow = 2)
33 annotate_figure(FIG,
34 top = text_grob("Vague versus Empirical Bayes: Beta-
  Bernoulli model", color = "black", face = "bold", size =
  14))

```

**FIGURE 1.1**

Vague versus Empirical Bayes: Bernoulli-Beta model.

*R code. Car claim, prior, likelihood and posterior
density plots*

```

1 # Prior, likelihood and posterior:
2 #Empirical Bayes Binomial-Beta model
3 dataNew <- data.frame(cbind(rep(lambda, 3),
4 c(EBPrior, PosteriorEB, Likelihood),
5 rep(1:3, each = 1000)))
6 #Data frame
7
8 colnames(dataNew) <- c("Lambda", "Density", "Factor")
9 dataNew$Factor <- factor(dataNew$Factor, levels=c("1", "3",
10 "2"), labels=c("Prior", "Likelihood", "Posterior"))
11
12 ggplot(data = dataNew, aes_string(x = "Lambda",
13 y = "Density", group = "Factor")) +
14 geom_line(aes(color = Factor)) +
15 xlab(TeX("$\\lambda$")) + ylab("Density") +
16 ggtitle("Prior, likelihood and posterior: Empirical Bayes
17 Poisson-Gamma model") +
18 guides(color=guide_legend(title="Information")) +
19 scale_color_manual(values = c("red", "yellow", "blue"))

```

**FIGURE 1.2**

Prior, likelihood and posterior: Bernoulli-Beta model.

R code. Car claim, predictive probabilities plots

```

1 # Predictive distributions
2 require(TailRank)
3 PredDen <- function(y, y0, a0, b0){
4   N <- length(y)
5   aN <- a0 + sum(y) # Posterior shape parameter
6   bN <- b0 + N - sum(y) # Posterior scale parameter
7   Pr <- aN/(aN+bN)
8   Probs <- dbinom(y0, 1, prob = Pr)
9   return(Probs)
10 }
11 y0 <- 0:1
12 PredVague <- PredDen(y = y, y0 = y0, a0 = a0, b0 = b0)
13 PredEB <- PredDen(y = y, y0 = y0, a0 = a0EB, b0 = b0EB)
14 dataPred <- as.data.frame(cbind(y0, PredVague, PredEB))
15 colnames(dataPred) <- c("y0", "PredictiveVague",
16   "PredictiveEB")
17 ggplot(data = dataPred) +
18   geom_point(aes(y0, PredictiveVague, color = "red")) +
19   xlab(TeX("$y_0$")) + ylab("Density") +
20   ggtitle("Predictive density: Vague and Empirical Bayes
21     priors") + geom_point(aes(y0, PredictiveEB, color = "
22     yellow")) +
23   guides(color = guide_legend(title="Prior")) +
24   scale_color_manual(labels = c("Vague", "Empirical Bayes"),
25     values = c("red", "yellow")) +
26   scale_x_continuous(breaks=seq(0,1,by=1))

```

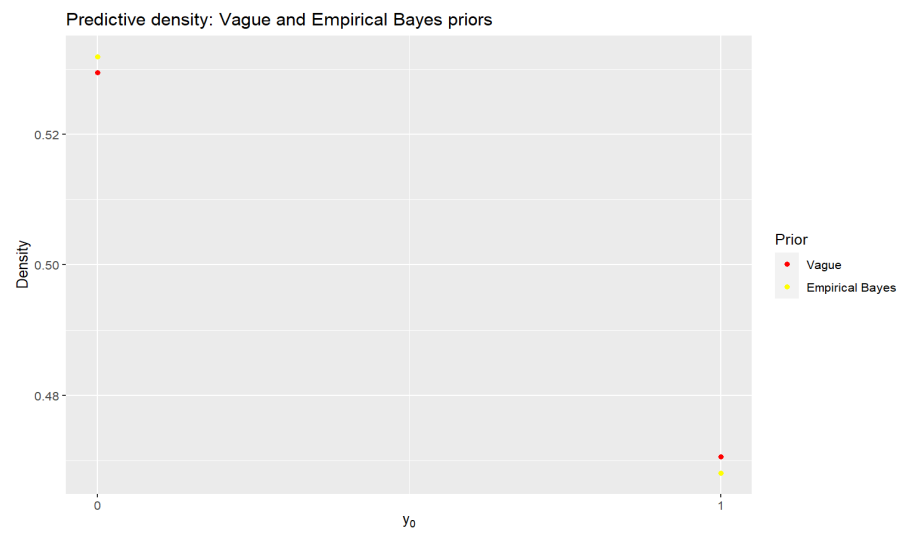


FIGURE 1.3
Predictive probabilities: Bernoulli-Beta model.

R code. Car claim, Bayesian model average

```

1 # Posterior odds: Vague vs Empirical Bayes
2 P012 <- exp(-LogMgLik(c(a0EB, b0EB), y = y))/exp(-LogMgLik(c
  (a0, b0), y = y))
3 PostProMEM <- P012/(1 + P012)
4 # Posterior model probability Empirical Bayes
5 PostProMEM
6 0.757
7 PostProbMV <- 1 - PostProMEM
8 # Posterior model probability vague prior
9 PostProbMV
10 0.242
11 # Bayesian model average (BMA)
12 PostMeanEB <- (a0EB + sum(y)) / (a0EB + b0EB + N)
13 # Posterior mean Empirical Bayes
14 PostMeanV <- (a0 + sum(y)) / (a0 + b0 + N)
15 # Posterior mean vague priors
16 BMAMean <- PostProMEM * PostMeanEB + PostProbMV * PostMeanV
17 # BMA posterior mean
18 PostVarEB <- (a0EB + sum(y))*(b0EB + N - sum(y)) / ((a0EB +
  b0EB + N)^2)*(a0EB + b0EB + N - 1)
19 # Posterior variance Empirical Bayes
20 PostVarV <- (a0 + sum(y))*(b0 + N - sum(y)) / ((a0 + b0 + N)
  ^2)*(a0 + b0 + N - 1)
21 # Posterior variance vague prior
22 BMAVar <- PostProMEM * PostVarEB + PostProbMV * PostVarV +
  PostProMEM * (PostMeanEB - BMAMean)^2 + PostProbMV * (
  PostMeanV - BMAMean)^2
23 # BMA posterior variance
24 # BMA: Predictive
25 BMAPred <- PostProMEM * PredEB + PostProbMV * PredVague
26 dataPredBMA <- as.data.frame(cbind(y0, BMAPred))
27 colnames(dataPredBMA) <- c("y0", "PredictiveBMA")
28 ggplot(data = dataPredBMA) +
29   geom_point(aes(y0, PredictiveBMA, color = "red")) +
30   xlab(TeX("$y_0$")) + ylab("Density") +
31   ggtitle("Predictive density: BMA") +
32   guides(color = guide_legend(title="BMA")) +
33   scale_color_manual(labels = c("Probability"), values = c("
  red")) + scale_x_continuous(breaks=seq(0,1,by=1))

```


R code. Car claim, Bayesian updating plots

```

1 # Bayesian updating
2 BayUp <- function(y, lambda, a0, b0){
3   N <- length(y)
4   aN <- a0 + sum(y)
5   # Posterior shape parameter
6   bN <- b0 + N - sum(y)
7   # Posterior scale parameter
8   p <- dbeta(lambda, shape1 = aN, shape2 = bN)
9   # Posterior density
10  return(list(Post = p, a0New = aN, b0New = bN))
11 }
12 PostUp <- NULL
13 for(i in 1:N){
14   if(i == 1){
15     PostUpi <- BayUp(y[i], lambda, a0 = 1, b0 = 1)}
16   else{
17     PostUpi <- BayUp(y[i], lambda,
18       a0 = PostUpi$a0New, b0 = PostUpi$b0New)
19   }
20   PostUp <- cbind(PostUp, PostUpi$Post)
21 }
22 DataUp <- data.frame(cbind(rep(lambda, 15), c(PostUp), rep
23   (1:15, each = 1000))) #Data frame
24 colnames(DataUp) <- c("Lambda", "Density", "Factor")
25 DataUp$Factor <- factor(DataUp$Factor, levels=c("1","2","3",
26   "4","5","6","7","8","9","10","11","12","13","14","15"),
27   labels=c("Iter_1","Iter_2","Iter_3","Iter_4","Iter_5","
28   Iter_6","Iter_7","Iter_8","Iter_9","Iter_10","Iter_11","
29   Iter_12","Iter_13","Iter_14","Iter_15"))
30 ggplot(data = DataUp, aes_string(x = "Lambda",
31   y = "Density", group = "Factor")) +
32   geom_line(aes(color = Factor)) +
33   xlab(TeX("$p$")) + ylab("Density") +
34   ggtitle("Bayesian updating:
35   Beta-Binomial model with vague prior") +
36   guides(color=guide_legend(title="Update"))

```

5. Show that given the loss function, $L(\theta, a) = |\theta - a|$, then the optimal decision rule minimizing the risk function, $a^*(\mathbf{y})$, is the median.

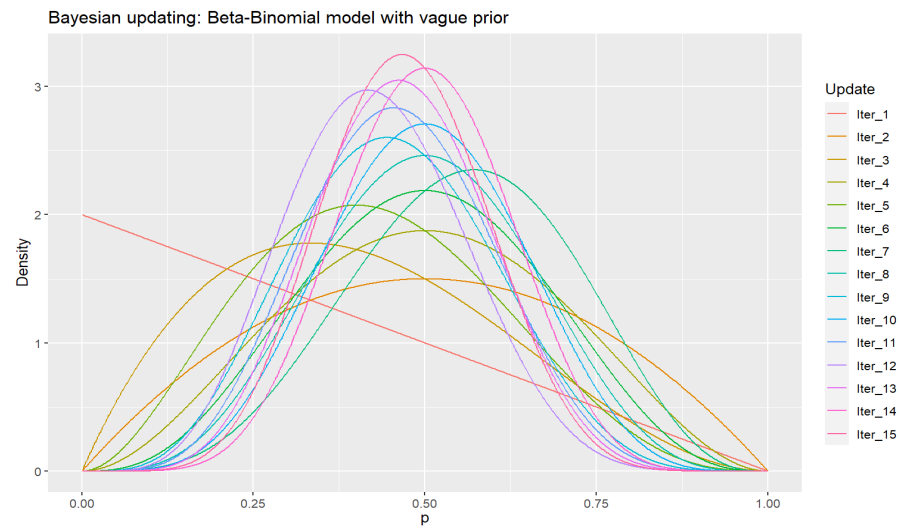
Answer

$\int_{\Theta} |\theta - a| \pi(\theta|\mathbf{y}) d\theta = \int_{-\infty}^a (a - \theta) \pi(\theta|\mathbf{y}) d\theta + \int_a^{\infty} (\theta - a) \pi(\theta|\mathbf{y}) d\theta$. Differentiating with respect to a , and equating to zero,

$$\int_{-\infty}^a \pi(\theta|\mathbf{y}) d\theta = \int_a^{\infty} \pi(\theta|\mathbf{y}) d\theta, \quad (1.13)$$

**FIGURE 1.4**

Predictive probabilities: Bernoulli-Beta Bayesian model average.

**FIGURE 1.5**

Predictive probabilities: Bernoulli-Beta Bayesian model updating.

then,

$$2 \int_{-\infty}^a \pi(\theta|\mathbf{y}) d\theta = \int_{-\infty}^{\infty} \pi(\theta|\mathbf{y}) d\theta = 1, \quad (1.14)$$

that is, $a^*(\mathbf{y})$ is the median.



2

Conceptual differences of the Bayesian and Frequentist approaches

2.1 Solutions of Exercises

1. Jeffreys-Lindley's paradox

The **Jeffreys-Lindley's paradox** [2, 4] is an apparent disagreement between the Bayesian and Frequentist frameworks to a hypothesis testing situation.

In particular, assume that in a city 49,581 boys and 48,870 girls have been born in 20 years. Assume that the male births is distributed Binomial with probability θ . We want to test the null hypothesis H_0 . $\theta = 0.5$ versus H_1 . $\theta \neq 0.5$.

- Show that the posterior model probability for the model under the null is approximately 0.95. Assume $\pi(H_0) = \pi(H_1) = 0.5$, and $\pi(\theta)$ equal to $U(0, 1)$ under H_1 .
- Show that the p -value for this hypothesis test is equal to 0.023 using the normal approximation, $Y \sim N(N \times \theta, N \times \theta \times (1 - \theta))$.

Answer

- The marginal likelihood under the null hypothesis is $p(y|H_0) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \approx 1.95 \times 10^{-4}$ given $\theta = 0.5$ under H_0 , $N = 49,581 + 48,870$ and $y = 49,581$. On the other hand, the marginal likelihood under the alternative hypothesis is

$$\begin{aligned}
p(y|H_1) &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta \\
&= \binom{n}{y} B(y+1, n-k+1) \\
&= \frac{\Gamma(N+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(y+1)\Gamma(N-y+1)}{\Gamma(N+2)} \\
&= \frac{N!}{(N+1)!} \\
&= \frac{1}{N+1} \\
&\approx 1.016 \times 10^{-5}.
\end{aligned}$$

Then, $PO_{01} = \frac{1.95 \times 10^{-4}}{1.016 \times 10^{-5}} = 19.19$, this implies that the posterior model probability under the null hypothesis is $\pi(H_0|y) = \frac{19.19}{1+19.19} = 0.95$.

- Under the null hypothesis,

$$\begin{aligned}
p &= 2 \int_{49,581}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} dy \\
&= 0.0235,
\end{aligned}$$

where $\mu = N \times \theta = 49,225.5$, and $\sigma^2 = N \times \theta \times (1-\theta) = 24,612.75$ under the null hypothesis ($\theta = 0.5$).

Observe that the posterior model probability supports the null hypothesis, whereas the p-value implies rejection of the null hypothesis using a 5% significance level.

Observe that actually this is not a paradox, as we are answering two different questions. The Bayes factor is comparing two models ($\theta = 0.5$ versus $\theta \sim U(0,1)$), whereas the p-value is checking the compatibility between $\theta = 0.5$ and the sample information. Despite that $\theta = 0.5$ is not compatible with sample information, it is better than the models assuming $\theta \sim U(0,1)$ as most of these values of θ are far away from the sample mean. Thus, the model under the null is a bad description of the data, but it is better than the model under the alternative hypothesis.¹

¹Observe that there are at least another two issues in this example. First, the prior under the alternative is non-informative, this implies problems for Bayes factors, and second, the prior under the alternative is positive at $\theta = 0.5$, which is the null ([3] propose non-local prior densities in Bayesian hypothesis tests to tackle these issues).

2. We want to test $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$ given $y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$.

Assume $\pi(H_0) = \pi(H_1) = 0.5$, and $\pi(\mu, \sigma) \propto 1/\sigma$ under the alternative hypothesis.

Show that

$$p(\mathbf{y}|\mathcal{M}_1) = \frac{\pi^{-N/2}}{2} \Gamma(N/2) 2^{N/2} \left(\frac{1}{\alpha_n \hat{\sigma}^2} \right)^{N/2} \left(\frac{N}{\alpha_n \hat{\sigma}^2} \right)^{-1/2} \frac{\Gamma(1/2) \Gamma(\alpha_n/2)}{\Gamma((\alpha_n+1)/2)} \text{ and}$$

$$p(\mathbf{y}|\mathcal{M}_0) = (2\pi)^{-N/2} \left[\frac{2}{\Gamma(N/2)} \left(\frac{N}{2} \frac{\sum_{i=1}^N (y_i - \mu_0)^2}{N} \right)^{N/2} \right]^{-1}. \text{ Then,}$$

$$PO_{01} = \frac{p(\mathbf{y}|\mathcal{M}_0)}{p(\mathbf{y}|\mathcal{M}_1)}$$

$$= \frac{\Gamma((\alpha_n + 1)/2)}{\Gamma(1/2) \Gamma(\alpha_n/2)} (\alpha_n \hat{\sigma}^2 / N)^{-1/2} \left[1 + \frac{(\mu_0 - \bar{y})^2}{\alpha_n \hat{\sigma}^2 / N} \right]^{-\left(\frac{\alpha_n + 1}{2}\right)},$$

where $\alpha_n = N - 1$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N - 1}$.

Find the relationship between the posterior odds and the classical test statistic for the null hypothesis.

Answer

$$\begin{aligned} p(\mathbf{y}|\mathcal{M}_1) &= \int_{-\infty}^{\infty} \int_0^{\infty} (2\pi)^{-N/2} \sigma^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \mu)^2 \right\} \frac{1}{\sigma} d\sigma d\mu \\ &= (2\pi)^{-N/2} \int_{-\infty}^{\infty} \int_0^{\infty} \sigma^{-(N+1)} \exp \left\{ -\frac{N}{2\sigma^2} \frac{\sum_{i=1}^N (y_i - \mu)^2}{N} \right\} d\sigma d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \int_{-\infty}^{\infty} \left[\sum_{i=1}^N (y_i - \mu)^2 \right]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \int_{-\infty}^{\infty} \left[\sum_{i=1}^N [(y_i - \bar{y}) - (\mu - \bar{y})]^2 \right]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \int_{-\infty}^{\infty} [\alpha_n \hat{\sigma}^2 + N(\mu - \bar{y})^2]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \left(\frac{\alpha_n \hat{\sigma}^2}{\alpha_n \hat{\sigma}^2} \right)^{-N/2} \int_{-\infty}^{\infty} [\alpha_n \hat{\sigma}^2 + N(\mu - \bar{y})^2]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} (\alpha_n \hat{\sigma}^2)^{-N/2} \int_{-\infty}^{\infty} \left[1 + \frac{N(\mu - \bar{y})^2}{\alpha_n \hat{\sigma}^2} \right]^{-N/2} d\mu \\ &= \frac{\pi^{-N/2}}{2} \Gamma(N/2) 2^{N/2} \left(\frac{1}{\alpha_n \hat{\sigma}^2} \right)^{N/2} \left(\frac{N}{\alpha_n \hat{\sigma}^2} \right)^{-1/2} \frac{\Gamma(1/2) \Gamma(\alpha_n/2)}{\Gamma((\alpha_n + 1)/2)}. \end{aligned}$$

The third line takes into account that the integral in the second line is the kernel of an inverted-gamma distribution, and the last line takes into account that the integral in the previous line is the kernel of a student's t distribution [7].

$$\begin{aligned}
p(\mathbf{y}|\mathcal{M}_0) &= \int_0^\infty (2\pi)^{-N/2} \sigma^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \mu_0)^2 \right\} \frac{1}{\sigma} d\sigma \\
&= (2\pi)^{-N/2} \int_0^\infty \sigma^{-(N+1)} \exp \left\{ -\frac{N}{2\sigma^2} \frac{\sum_{i=1}^N (y_i - \mu_0)^2}{N} \right\} d\sigma \\
&= (2\pi)^{-N/2} \left[\frac{2}{\Gamma(N/2)} \left(\frac{N}{2} \frac{\sum_{i=1}^N (y_i - \mu_0)^2}{N} \right)^{N/2} \right]^{-1}.
\end{aligned}$$

The third line takes into account that the integral in the second line is the kernel of an inverted-gamma distribution [7].

Given these results is easy to get PO_{01} .

In addition,

$$\begin{aligned}
PO_{01} &= \frac{\Gamma((\alpha_n + 1)/2)}{\Gamma(1/2)\Gamma(\alpha_N/2)} (\alpha_n \hat{\sigma}^2/N)^{-1/2} \left[1 + \frac{(\mu_0 - \bar{y})^2}{\alpha_n \hat{\sigma}^2/N} \right]^{-\left(\frac{\alpha_n+1}{2}\right)} \\
&= \frac{\Gamma((\alpha_n + 1)/2)}{\Gamma(1/2)\Gamma(\alpha_N/2)} (\alpha_n \hat{\sigma}^2/N)^{-1/2} \left[1 + \frac{1}{\alpha_n} \left(\frac{\mu_0 - \bar{y}}{\hat{\sigma}/\sqrt{N}} \right)^2 \right]^{-\left(\frac{\alpha_n+1}{2}\right)} \\
&= \frac{\Gamma((\alpha_n + 1)/2)}{\Gamma(1/2)\Gamma(\alpha_N/2)} (\alpha_n \hat{\sigma}^2/N)^{-1/2} \left[1 + \frac{1}{\alpha_n} t^2 \right]^{-\left(\frac{\alpha_n+1}{2}\right)},
\end{aligned}$$

where $t = \frac{\bar{y} - \mu_0}{\hat{\sigma}/\sqrt{N}}$ is the classical statistical test. Then, as t increases then the PO_{01} decreases, both indicating support against the null hypothesis H_0 . $\mu = \mu_0$. However, there are other terms affecting the posterior odds, then, there is no necessary agreement between the classical test statistic and the posterior odds.

3. Math test continues

Using the setting of the **Example: Math test** in subsection 2.6.1 in the book, test H_0 . $\mu = \mu_0$ vs H_1 . $\mu \neq \mu_0$ where $\mu_0 = \{100, 100.5, 101, 101.5, 102\}$.

- What is the p -value for these hypothesis tests?
- Find the posterior model probability of the null model for each μ_0 .

R code. Example: Math test

```
1 N <- 50 # Sample size
2 y_bar <- 102 # Sample mean
3 s2 <- 10 # Sample variance
4 alpha <- N - 1
5 serror <- (s2/N)^0.5
6 y.H0 <- c(100, 100.5, 101, 101.5, 102)
7 test <- (y.H0 - y_bar)/serror
8 pval <- 2*pt(test, alpha)
9 pval
10 0.0000459 0.0015431 0.0299338 0.2690040 1
11 # p-values
12 P001 <- (gamma(N/2)*((N-1)*serror^2)^(-0.5)*(1+test^2/alpha)
13         ^(-N/2))/(gamma(1/2)*gamma((N-1)/2))
14 P001/(1+P001)
15 0.0001705 0.0050345 0.0725330 0.3210223 0.4702050
16 # Posterior model probability of the null hypothesis.
```



3

Objective and subjective Bayesian approaches

3.1 Solutions of Exercises

1. Elicitation ...

R code. Example: Math test

```
1 N <- 50 # Sample size
2 y_bar <- 102 # Sample mean
3 s2 <- 10 # Sample variance
4 alpha <- N - 1
5 serror <- (s2/N)^0.5
6 y.H0 <- c(100, 100.5, 101, 101.5, 102)
7 test <- (y.H0 - y_bar)/serror
8 pval <- 2*pt(test, alpha)
9 pval
10 0.0000459 0.0015431 0.0299338 0.2690040 1
11 # p-values
12 P001 <- (gamma(N/2)*((N-1)*serror^2)^(-0.5)*(1+test^2/alpha)
13         ^(-N/2))/(gamma(1/2)*gamma((N-1)/2))
14 P001/(1+P001)
15 0.0001705 0.0050345 0.0725330 0.3210223 0.4702050
16 # Posterior model probability of the null hypothesis.
```



4

Cornerstone models: Conjugate families

4.1 Solutions of Exercises

1. Write in the canonical form the distribution of the Bernoulli example, and find the mean and variance of the sufficient statistic.

Answer

Given $p(\mathbf{y}|\theta) = (1-\theta)^N \exp \left\{ \sum_{i=1}^N y_i \log \left(\frac{\theta}{1-\theta} \right) \right\}$ where $\eta = \log \frac{\theta}{1-\theta}$ which implies $\theta = \frac{\exp(\eta)}{1+\exp(\eta)}$, then $p(\mathbf{y}|\theta) = \exp \left\{ \sum_{i=1}^N y_i \eta - N \log(1 + \exp(\eta)) \right\}$. Thus $B(\eta) = N \log(1 + \exp(\eta))$, $\nabla(B(\eta)) = N \frac{\exp(\eta)}{1+\exp(\eta)} = N\theta$ and $\nabla^2(B(\eta)) = N \left\{ \frac{\exp(\eta)(1+\exp(\eta))}{(1+\exp(\eta))^2} - \frac{\exp(\eta)\exp(\eta)}{(1+\exp(\eta))^2} \right\} = N\theta(1-\theta)$.

2. Given a random sample $\mathbf{y} = [y_1, y_2, \dots, y_N]^\top$ from N binomial experiments each having known size n_i and same unknown probability θ . Show that $p(\mathbf{y}|\theta)$ is in the exponential family, and find the posterior distribution, the marginal likelihood and the predictive distribution of the binomial-beta model assuming the number of trials is known.

Answer

The density function is

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^N \binom{n_i}{y_i} \theta^{y_i} (1-\theta)^{n_i-y_i} \\ &= \prod_{i=1}^N \binom{n_i}{y_i} \theta^{\sum_{i=1}^N y_i} (1-\theta)^{\sum_{i=1}^N n_i - \sum_{i=1}^N y_i} \\ &= \prod_{i=1}^N \binom{n_i}{y_i} \exp \left\{ \sum_{i=1}^N y_i \log \left(\frac{\theta}{1-\theta} \right) + \sum_{i=1}^N n_i \log(1-\theta) \right\} \\ &= \prod_{i=1}^N \binom{n_i}{y_i} (1-\theta)^{\sum_{i=1}^N n_i} \exp \left\{ \sum_{i=1}^N y_i \log \left(\frac{\theta}{1-\theta} \right) \right\}, \end{aligned}$$

Observe that $\sum_{i=1}^N n_i$ is the total sample size of Bernoulli experiments.

Using Theorem 1 in Chapter 4, the prior distribution is

$$\begin{aligned}\pi(\theta) &\propto (1-\theta)^{B_0} \exp \left\{ a_0 \log \left(\frac{\theta}{1-\theta} \right) \right\} \\ &= \theta^{a_0} (1-\theta)^{B_0-a_0} \\ &= \theta^{\alpha_0-1} (1-\theta)^{\beta_0-1},\end{aligned}$$

where $\alpha_0 = a_0 + 1$ and $\beta_0 = B_0 - a_0 + 1$. This is the kernel of a beta distribution. Thus, the posterior distribution is

$$\begin{aligned}\pi(\theta|\mathbf{y}) &\propto \theta^{\alpha_0-1} (1-\theta)^{\beta_0-1} \times \theta^{\sum_{i=1}^N y_i} (1-\theta)^{\sum_{i=1}^N n_i - \sum_{i=1}^N y_i} \\ &= \theta^{\alpha_0 + \sum_{i=1}^N y_i - 1} (1-\theta)^{\beta_0 + \sum_{i=1}^N n_i - \sum_{i=1}^N y_i - 1} \\ &= \theta^{\alpha_n - 1} (1-\theta)^{\beta_n - 1},\end{aligned}$$

where $\alpha_n = \alpha_0 + \sum_{i=1}^N y_i$ and $\beta_n = \beta_0 + \sum_{i=1}^N n_i - \sum_{i=1}^N y_i$.

The marginal likelihood is

$$\begin{aligned}p(\mathbf{y}) &= \int_0^1 \frac{\theta^{\alpha_0-1} (1-\theta)^{\beta_0-1}}{B(\alpha_0, \beta_0)} \times \prod_{i=1}^N \binom{n_i}{y_i} \theta^{\sum_{i=1}^N y_i} (1-\theta)^{\sum_{i=1}^N n_i - \sum_{i=1}^N y_i} d\theta \\ &= \frac{\prod_{i=1}^N \binom{n_i}{y_i}}{B(\alpha_0, \beta_0)} \int_0^1 \theta^{\alpha_0 + \sum_{i=1}^N y_i - 1} (1-\theta)^{\beta_0 + \sum_{i=1}^N n_i - \sum_{i=1}^N y_i - 1} d\theta \\ &= \frac{\prod_{i=1}^N \binom{n_i}{y_i} B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)}.\end{aligned}$$

The third line due to having the kernel of a Beta distribution.

Finally, the predictive distribution is

$$\begin{aligned}p(Y_0|\mathbf{y}) &= \int_0^1 \binom{n_{y_0}}{y_0} \theta^{y_0} (1-\theta)^{n_{y_0}-y_0} \frac{\theta^{\alpha_n-1} (1-\theta)^{\beta_n-1}}{B(\alpha_n, \beta_n)} d\theta \\ &= \frac{\binom{n_{y_0}}{y_0}}{B(\alpha_n, \beta_n)} \int_0^1 \theta^{\alpha_n + y_0 - 1} (1-\theta)^{\beta_n + n_{y_0} - y_0 - 1} d\theta \\ &= \binom{n_{y_0}}{y_0} \frac{B(\alpha_n + y_0, \beta_n + n_{y_0} - y_0)}{B(\alpha_n, \beta_n)},\end{aligned}$$

where n_{y_0} is the known size associated with y_0 , and the last line due to having the kernel of a beta distribution. The predictive is a *beta-binomial distribution*.

3. Given a random sample $\mathbf{y} = [y_1, y_2, \dots, y_N]^\top$ from a *exponential distribution*. Show that $p(\mathbf{y}|\lambda)$ is in the exponential family, and find the posterior distribution, marginal likelihood and predictive distribution of the exponential-gamma model.

Answer

We see that the exponential distribution belongs to the exponential family as $p(\mathbf{y}|\lambda) = \prod_{i=1}^N \lambda \exp(-\lambda y_i) = \lambda^N \exp(-\lambda \sum_{i=1}^N y_i)$.

Using the gamma distribution in the rate parametrization, we see that $\pi(\lambda|\mathbf{y}) \propto \lambda^{\alpha_0-1} \exp(-\lambda\beta_0) \times \lambda^N \exp(-\lambda \sum_{i=1}^N y_i) = \lambda^{\alpha_0+N-1} \exp(-\lambda(\beta_0 + \sum_{i=1}^N y_i))$. This is the kernel of a gamma distribution, that is, $\lambda|\mathbf{y} \sim G(\alpha_n, \beta_n)$ where $\alpha_n = \alpha_0 + N$ and $\beta_n = \beta_0 + \sum_{i=1}^N y_i$.

The marginal likelihood is

$$\begin{aligned} p(\mathbf{y}) &= \int_0^\infty \lambda^N \exp\left\{-\lambda \sum_{i=1}^N y_i\right\} \lambda^{\alpha_0-1} \exp\{-\beta_0 \lambda\} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} d\lambda \\ &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \int_0^\infty \lambda^{\alpha_0+N-1} \exp\left\{-\lambda \left(\beta_0 + \sum_{i=1}^N y_i\right)\right\} d\lambda \\ &= \frac{\beta_0^{\alpha_0} \Gamma(\alpha_n)}{\Gamma(\alpha_0) \beta_n^{\alpha_n}}. \end{aligned}$$

Finally, the predictive distribution is

$$\begin{aligned} p(Y_0|\mathbf{y}) &= \int_0^\infty \lambda \exp\{-\lambda y_0\} \lambda^{\alpha_n-1} \exp\{-\beta_n \lambda\} \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} d\lambda \\ &= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \int_0^\infty \lambda^{\alpha_n+1-1} \exp\{-\lambda(\beta_n + y_0)\} d\lambda \\ &= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \times \frac{\Gamma(\alpha_n + 1)}{(\beta_n + y_0)^{\alpha_n+1}} \\ &= \frac{\alpha_n \beta_n^{\alpha_n}}{(\beta_n + y_0)^{\alpha_n+1}}. \end{aligned}$$

This is a *Lomax distribution*.

4. Given $\mathbf{y} \sim N_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, that is, a *multivariate normal distribution* show that $p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is in the exponential family.

Answer

$$\begin{aligned}
p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-N/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\} \\
&= (2\pi)^{-N/2} \exp \left\{ -\frac{1}{2} (\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} - 2\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \log(|\boldsymbol{\Sigma}|)) \right\} \\
&= (2\pi)^{-N/2} \exp \left\{ -\frac{1}{2} (tr \{ \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \} - 2\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \log(|\boldsymbol{\Sigma}|)) \right\} \\
&= (2\pi)^{-N/2} \exp \left\{ -\frac{1}{2} (vec(\mathbf{y}\mathbf{y}^\top)^\top vec(\boldsymbol{\Sigma}^{-1}) - 2\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \log(|\boldsymbol{\Sigma}|)) \right\},
\end{aligned}$$

where tr and vec are the trace and vectorization operators, respectively.

Then, $h(\mathbf{y}) = (2\pi)^{-N/2}$, $\eta(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = [\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \quad vec(\boldsymbol{\Sigma}^{-1})]$, $T(\mathbf{y}) = [\mathbf{y} \quad \frac{1}{2} vec(\mathbf{y}\mathbf{y}^\top)]$ and $C(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp \left\{ -\frac{1}{2N} (\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \log(|\boldsymbol{\Sigma}|)) \right\}$.

5. Find the marginal likelihood in the normal/inverse-Wishart model.

Answer

$$\begin{aligned}
p(\mathbf{Y}) &= \int_{\mathcal{R}^p} \int_{\mathcal{S}} (2\pi)^{-pN/2} |\boldsymbol{\Sigma}|^{-N/2} \exp \left\{ -\frac{1}{2} tr[(\mathbf{S} + N(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top) \boldsymbol{\Sigma}^{-1}] \right\} \\
&\quad \times (2\pi)^{-p/2} \beta_0^{p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{\beta_0}{2} tr[(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}] \right\} \\
&\quad \times |\boldsymbol{\Sigma}|^{-(\alpha_0 + p + 1)/2} \frac{2^{-\alpha_0 p/2} |\boldsymbol{\Psi}_0|^{\alpha_0/2}}{\Gamma_p(\alpha_0/2)} \exp \left\{ -\frac{1}{2} tr(\boldsymbol{\Psi}_0 \boldsymbol{\Sigma}^{-1}) \right\} d\boldsymbol{\Sigma} d\boldsymbol{\mu} \\
&= \frac{(2\pi)^{-frac{1}{2}(pN+p)} |\boldsymbol{\Psi}_0|^{\alpha_0/2} \beta_0^{p/2} 2^{-\alpha_0 p/2}}{\Gamma_p(\alpha_0/2)} \int_{\mathcal{R}^p} \int_{\mathcal{S}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}(N+1+\alpha_0+p+1)} \\
&\quad \times \exp \left\{ -\frac{1}{2} tr[(\mathbf{S} + N(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top + \beta_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top + \boldsymbol{\Psi}_0) \boldsymbol{\Sigma}^{-1}] \right\} d\boldsymbol{\Sigma} d\boldsymbol{\mu}.
\end{aligned}$$

We have in the integral the kernel of an Inverse-Wishart distribution, then

$$\begin{aligned}
p(\mathbf{Y}) &= \frac{\Gamma_p\left(\frac{N+1+\alpha_0}{2}\right) |\Psi_0|^{\alpha_0/2} \beta_0^{p/2}}{\Gamma_p(\alpha_0/2) \pi^{p(N+1)/2}} \\
&\quad \times \int_{\mathcal{R}^p} |\mathbf{S} + \Psi_0 + (N + \beta_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top \\
&\quad + N\beta_0/(N + \beta_0)(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^\top| d\boldsymbol{\mu} \\
&= \frac{\Gamma_p\left(\frac{N+1+\alpha_0}{2}\right) |\Psi_0|^{\alpha_0/2} \beta_0^{p/2}}{\Gamma_p(\alpha_0/2) \pi^{p(N+1)/2}} \\
&\quad \times \int_{\mathcal{R}^p} |\Psi_n| |1 + \beta_n(\boldsymbol{\mu} - \boldsymbol{\mu}_n) \Psi_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top|^{-\frac{1}{2}(\alpha_n+1)} d\boldsymbol{\mu} \\
&= \frac{\Gamma_p\left(\frac{\alpha_n+1}{2}\right) |\Psi_0|^{\alpha_0/2} \beta_0^{p/2}}{\Gamma_p(\alpha_0/2) \pi^{p(N+1)/2}} |\Psi_n|^{-\frac{1}{2}(\alpha_n+1)} \\
&\quad \times \int_{\mathcal{R}^p} [1 + \beta_n(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top \Psi_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)]^{-\frac{1}{2}(\alpha_n+1)} d\boldsymbol{\mu}.
\end{aligned}$$

The last equality uses the definition of Ψ_n , β_n and α_n , and the Sylvester's determinant theorem. Observe that we have the kernel of a multivariate t distribution [5]. Then,

$$\begin{aligned}
p(\mathbf{Y}) &= \frac{\Gamma_p\left(\frac{\alpha_n+1}{2}\right) |\Psi_0|^{\alpha_0/2} \beta_0^{p/2}}{\Gamma_p(\alpha_0/2) \pi^{p(N+1)/2}} |\Psi_n|^{-\frac{1}{2}(\alpha_n+1)} \\
&\quad \times \int_{\mathcal{R}^p} \left[1 + \frac{1}{\alpha_n + 1 - p} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top \left(\frac{\Psi_n}{\beta_n(\alpha_n + 1 - p)} \right)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]^{-\frac{1}{2}(\alpha_n+1-p+p)} d\boldsymbol{\mu} \\
&= \frac{\Gamma_p\left(\frac{\alpha_n+1}{2}\right) \Gamma_p\left(\frac{\alpha_n+1-p}{2}\right) |\Psi_0|^{\alpha_0/2} \beta_0^{p/2} (\alpha_n + 1 - p)^{p/2} \pi^{p/2} |\Psi_n|^{-\frac{1}{2}(\alpha_n+1)}}{\Gamma_p(\alpha_0/2) \pi^{p(N+1)/2} \Gamma_p\left(\frac{\alpha_n+1-p+p}{2}\right) \left(\frac{\Psi_n}{\alpha_n+1-p} \right)^{-1/2}} \\
&= \frac{\Gamma_p\left(\frac{v_n}{2}\right) |\Psi_0|^{\alpha_0/2}}{\Gamma_p\left(\frac{\alpha_0}{2}\right) |\Psi_n|^{\alpha_n/2}} \left(\frac{\beta_0}{\beta_n} \right)^{p/2} (2\pi)^{-Np/2},
\end{aligned}$$

where $v_n = \alpha_n + 1 - p$.

6. Find the posterior predictive distribution in the normal/inverse-Wishart model, and show that $\mathbf{Y}_0|\mathbf{Y} \sim T_{N_0,M}(\alpha_n - M + 1, \mathbf{X}_0\mathbf{B}_n, \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top, \Psi_n)$ in the multivariate regression linear model.

Answer

$$\begin{aligned}
p(\mathbf{Y}_0|\mathbf{Y}) &\propto \int_{\mathcal{R}^p} \int_S |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \text{tr}[(\mathbf{y}_0 - \boldsymbol{\mu})(\mathbf{y}_0 - \boldsymbol{\mu})^\top \Sigma^{-1}] \right\} \\
&\quad \times |\Sigma|^{-1/2} \exp \left\{ -\frac{\beta_n}{2} \text{tr}[(\boldsymbol{\mu} - \boldsymbol{\mu}_n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top \Sigma^{-1}] \right\} \\
&\quad \times |\Sigma|^{-(\alpha_n + p + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Psi_n \Sigma^{-1}) \right\} d\Sigma d\boldsymbol{\mu} \\
&\propto \int_{\mathcal{R}^p} |(\mathbf{y}_0 - \boldsymbol{\mu})(\mathbf{y}_0 - \boldsymbol{\mu})^\top + (\boldsymbol{\mu} - \boldsymbol{\mu}_n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top + \Psi_n|^{-(\alpha_n + 2)/2} d\boldsymbol{\mu}.
\end{aligned}$$

The last equality uses that there is the kernel of an Inverse Wishart distribution.

Taking into account that

$$\begin{aligned}
(\mathbf{y}_0 - \boldsymbol{\mu})(\mathbf{y}_0 - \boldsymbol{\mu})^\top + (\boldsymbol{\mu} - \boldsymbol{\mu}_n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top &= (1 + \beta_n) \left(\boldsymbol{\mu} - \frac{(\mathbf{y}_0 + \beta_n \boldsymbol{\mu}_n)}{1 + \beta_n} \right) \left(\boldsymbol{\mu} - \frac{(\mathbf{y}_0 + \beta_n \boldsymbol{\mu}_n)}{1 + \beta_n} \right)^\top \\
&\quad + \frac{\beta_n}{1 + \beta_n} (\mathbf{y}_0 - \boldsymbol{\mu}_n)(\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top.
\end{aligned}$$

Then,

$$\begin{aligned}
p(\mathbf{Y}_0|\mathbf{Y}) &\propto \int_{\mathcal{R}^p} |(\mathbf{y}_0 - \boldsymbol{\mu})(\mathbf{y}_0 - \boldsymbol{\mu})^\top + (\boldsymbol{\mu} - \boldsymbol{\mu}_n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top + \Psi_n|^{-(\alpha_n + 2)/2} d\boldsymbol{\mu} \\
&= \int_{\mathcal{R}^p} \left| (1 + \beta_n) \left(\boldsymbol{\mu} - \frac{(\mathbf{y}_0 + \beta_n \boldsymbol{\mu}_n)}{1 + \beta_n} \right) \left(\boldsymbol{\mu} - \frac{(\mathbf{y}_0 + \beta_n \boldsymbol{\mu}_n)}{1 + \beta_n} \right)^\top \right. \\
&\quad \left. + \frac{\beta_n}{1 + \beta_n} (\mathbf{y}_0 - \boldsymbol{\mu}_n)(\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top + \Psi_n \right|^{-(\alpha_n + 2)/2} d\boldsymbol{\mu} \\
&= \int_{\mathcal{R}^p} \left| \underbrace{\Psi_n + \frac{\beta_n}{1 + \beta_n} (\mathbf{y}_0 - \boldsymbol{\mu}_n)(\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top}_{\Lambda_n} \right| \\
&\quad \left| 1 + (1 + \beta_n) \left(\boldsymbol{\mu} - \frac{(\mathbf{y}_0 + \beta_n \boldsymbol{\mu}_n)}{1 + \beta_n} \right)^\top \frac{1}{\alpha_n + 2 - p} \left(\frac{\Lambda_n}{\alpha_n + 2 - p} \right)^{-1} \left(\boldsymbol{\mu} - \frac{(\mathbf{y}_0 + \beta_n \boldsymbol{\mu}_n)}{1 + \beta_n} \right) \right|^{-(\alpha_n + 2 - p + p)/2} d\boldsymbol{\mu} \\
&\propto \left| \Psi_n + \frac{\beta_n}{1 + \beta_n} (\mathbf{y}_0 - \boldsymbol{\mu}_n)(\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top \right|^{-(\alpha_n + 2)/2} \\
&\quad \times \left| \Psi_n + \frac{\beta_n}{1 + \beta_n} (\mathbf{y}_0 - \boldsymbol{\mu}_n)(\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top \right|^{1/2} \\
&= \left| \Psi_n + \frac{\beta_n}{1 + \beta_n} (\mathbf{y}_0 - \boldsymbol{\mu}_n)(\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top \right|^{-(\alpha_n + 1)/2} \\
&\propto \left[1 + (\mathbf{y}_0 - \boldsymbol{\mu}_n)^\top \frac{1}{\alpha_n + 1 - p} \left(\frac{\Psi_n (1 + \beta_n)}{(\alpha_n + 1 - p) \beta_n} \right)^{-1} (\mathbf{y}_0 - \boldsymbol{\mu}_n) \right]^{-(\alpha_n + 1 - p + p)}.
\end{aligned}$$

The second equality and last line use the Sylvester's determinant theorem, and the second equality uses that there is the kernel of a multivariate t distribution.

Then, we have that the predictive distribution is a multivariate t distribution centered at $\boldsymbol{\mu}_n$, $\alpha_n + 1 - p$ degrees of freedom, and scale matrix $\frac{\boldsymbol{\Psi}_n(1+\beta_n)}{(\alpha_n+1-p)\beta_n}$.

To show the second statement, let's start by the definition of the predictive density to show that $\mathbf{Y}_0|\mathbf{Y} \sim T_{N_0,M}(\alpha_n - M + 1, \mathbf{X}_0\mathbf{B}_n, \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top, \boldsymbol{\Psi}_n)$.

$$\begin{aligned} \pi(\mathbf{Y}_0|\mathbf{Y}) &\propto \int_{\mathcal{S}} \int_{\mathcal{B}} \left\{ |\boldsymbol{\Sigma}|^{-N_0/2} \exp \left\{ -\frac{1}{2} \text{tr}[(\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B})^\top (\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B}) \boldsymbol{\Sigma}^{-1}] \right\} \right. \\ &\quad \times |\boldsymbol{\Sigma}|^{-K/2} \exp \left\{ -\frac{1}{2} \text{tr}[(\mathbf{B} - \mathbf{B}_n)^\top \mathbf{V}_n^{-1} (\mathbf{B} - \mathbf{B}_n) \boldsymbol{\Sigma}^{-1}] \right\} \\ &\quad \times |\boldsymbol{\Sigma}|^{-(\alpha_n+M+1)/2} \exp \left\{ -\frac{1}{2} \text{tr}[\boldsymbol{\Psi}_n \boldsymbol{\Sigma}^{-1}] \right\} \Big\} d\mathbf{B} d\boldsymbol{\Sigma} \\ &= \int_{\mathcal{S}} \int_{\mathcal{B}} \left\{ |\boldsymbol{\Sigma}|^{-(N_0+K+\alpha_n+M+1)/2} \exp \left\{ -\frac{1}{2} \text{tr} [((\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B})^\top (\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B}) \right. \right. \\ &\quad \left. \left. + (\mathbf{B} - \mathbf{B}_n)^\top \mathbf{V}_n^{-1} (\mathbf{B} - \mathbf{B}_n) + \boldsymbol{\Psi}_n) \boldsymbol{\Sigma}^{-1}] \right\} \right\} d\mathbf{B} d\boldsymbol{\Sigma}. \end{aligned}$$

Setting $\mathbf{M} = (\mathbf{X}_0^\top \mathbf{X}_0 + \mathbf{V}_n^{-1})$, and $\mathbf{B}_* = \mathbf{M}^{-1}(\mathbf{V}_n \mathbf{B}_n + \mathbf{X}_0^\top \mathbf{Y}_0)$, we have that $(\mathbf{B} - \mathbf{B}_*)^\top \mathbf{M} (\mathbf{B} - \mathbf{B}_*) + \mathbf{B}_n^\top \mathbf{V}_n^{-1} \mathbf{B}_n + \mathbf{Y}_0^\top \mathbf{Y}_0 - \mathbf{B}_*^\top \mathbf{M} \mathbf{B}_* = (\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B})^\top (\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B}) + (\mathbf{B} - \mathbf{B}_n)^\top \mathbf{V}_n^{-1} (\mathbf{B} - \mathbf{B}_n)$. Then,

$$\begin{aligned} \pi(\mathbf{Y}_0|\mathbf{Y}) &\propto \int_{\mathcal{S}} |\boldsymbol{\Sigma}|^{-(N_0+K+\alpha_n+M+1)/2} \\ &\quad \times \exp \left\{ -\frac{1}{2} \text{tr}[(\boldsymbol{\Psi}_n + \mathbf{B}_n^\top \mathbf{V}_n^{-1} \mathbf{B}_n + \mathbf{Y}_0^\top \mathbf{Y}_0 - \mathbf{B}_*^\top \mathbf{M} \mathbf{B}_*) \boldsymbol{\Sigma}^{-1}] \right\} \\ &\quad \times \int_{\mathcal{B}} \exp \left\{ -\frac{1}{2} \text{tr}[(\mathbf{B} - \mathbf{B}_*)^\top \mathbf{M} (\mathbf{B} - \mathbf{B}_*) \boldsymbol{\Sigma}^{-1}] \right\} d\mathbf{B} d\boldsymbol{\Sigma}. \end{aligned}$$

The latter is the kernel of a matrix normal distribution, thus

$$\begin{aligned} \pi(\mathbf{Y}_0|\mathbf{Y}) &\propto \int_{\mathcal{S}} |\boldsymbol{\Sigma}|^{-(N_0+\alpha_n+M+1)/2} \\ &\quad \times \exp \left\{ -\frac{1}{2} \text{tr}[(\boldsymbol{\Psi}_n + \mathbf{B}_n^\top \mathbf{V}_n^{-1} \mathbf{B}_n + \mathbf{Y}_0^\top \mathbf{Y}_0 - \mathbf{B}_*^\top \mathbf{M} \mathbf{B}_*) \boldsymbol{\Sigma}^{-1}] \right\} d\boldsymbol{\Sigma} \end{aligned}$$

This is the kernel of an inverse-Wishart distribution, then

$$\pi(\mathbf{Y}_0|\mathbf{Y}) \propto |\boldsymbol{\Psi}_n + \mathbf{B}_n^\top \mathbf{V}_n^{-1} \mathbf{B}_n + \mathbf{Y}_0^\top \mathbf{Y}_0 - \mathbf{B}_*^\top \mathbf{M} \mathbf{B}_*|^{-(N_0+\alpha_n)/2}.$$

Setting $\mathbf{C}^{-1} = \mathbf{I}_{N_0} + \mathbf{X}_0 \mathbf{V}_n \mathbf{X}_0^\top$ such that $\mathbf{C} = \mathbf{I}_{N_0} - \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0 + \mathbf{V}_n^{-1})^{-1} \mathbf{X}_0^\top$ (see footnote 4 in Chapter 4), then $\mathbf{B}_n^\top \mathbf{V}_n^{-1} \mathbf{B}_n + \mathbf{Y}_0^\top \mathbf{Y}_0 - \mathbf{B}_*^\top \mathbf{M} \mathbf{B}_* = (\mathbf{Y}_0 - \mathbf{X}_0 \mathbf{B}_n)^\top \mathbf{C} (\mathbf{Y}_0 - \mathbf{X}_0 \mathbf{B}_n)$. This is done following exactly same procedure as deducing the predictive distribution in the linear regression model in the book. Thus,

$$\begin{aligned} \pi(\mathbf{Y}_0 | \mathbf{Y}) &\propto |\boldsymbol{\Psi}_n + (\mathbf{Y}_0 - \mathbf{X}_0 \mathbf{B}_n)^\top \mathbf{C} (\mathbf{Y}_0 - \mathbf{X}_0 \mathbf{B}_n)|^{-(N_0 + \alpha_n)/2} \\ &\propto |\mathbf{I}_{N_0} + \mathbf{C} (\mathbf{Y}_0 - \mathbf{X}_0 \mathbf{B}_n) \boldsymbol{\Psi}^{-1} (\mathbf{Y}_0 - \mathbf{X}_0 \mathbf{B}_n)^\top|^{-(\alpha_n + 1 - M + N_0 + M - 1)/2}. \end{aligned}$$

The second proportionality follows from the Sylvester's theorem. Observe that this is the kernel of a matrix t distribution with $\alpha_n + 1 - M$ degrees of freedom, location $\mathbf{X}_0 \mathbf{B}_n$ and scale matrices $\boldsymbol{\Psi}_n$ and $\mathbf{C}^{-1} = \mathbf{I}_{N_0} + \mathbf{X}_0 \mathbf{V}_n \mathbf{X}_0^\top$.

7. Show that $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top ((\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0)^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ in the linear regression model, and that $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \mathbf{S} + (\hat{\mathbf{B}} - \mathbf{B}_0)^\top \mathbf{V}_n (\hat{\mathbf{B}} - \mathbf{B}_0)$ in the linear multivariate regression model.

Answer

Taking into account that

$$\begin{aligned} \delta^* &= \delta_0 + \mathbf{y}^\top \mathbf{y} + \boldsymbol{\beta}_0^\top \mathbf{B}_0^{-1} \boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top \mathbf{B}_n^{-1} \boldsymbol{\beta}_n \\ &= \delta_0 + \mathbf{y}^\top \mathbf{y} + \boldsymbol{\beta}_0^\top \mathbf{B}_0^{-1} \boldsymbol{\beta}_0 - (\mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}})^\top \mathbf{B}_n (\mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}) \\ &= \delta_0 + \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\top (\mathbf{B}_0^{-1} - \mathbf{B}_0^{-1} \mathbf{B}_n \mathbf{B}_0^{-1}) \boldsymbol{\beta}_0 \\ &\quad - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \delta_0 + \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} \\ &\quad - 2\hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\top (\mathbf{B}_0^{-1} - \mathbf{B}_0^{-1} \mathbf{B}_n \mathbf{B}_0^{-1}) \boldsymbol{\beta}_0. \end{aligned}$$

Observe that

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \mathbf{y}^\top \mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y} + \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \mathbf{y}^\top \mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top (\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\mu}}) + \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}, \end{aligned}$$

where $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\mu}}$, and $\mathbf{X}^\top \hat{\boldsymbol{\mu}} = 0$.

The following matrix identities are useful [6]:

$$(\mathbf{D} + \mathbf{E})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1} (\mathbf{D}^{-1} + \mathbf{E}^{-1})^{-1} \mathbf{D}^{-1},$$

and

$$(\mathbf{D} + \mathbf{E})^{-1} = \mathbf{D}^{-1} (\mathbf{E}^{-1} + \mathbf{D}^{-1}) \mathbf{E}^{-1}.$$

Using these identities,

$$\begin{aligned} [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} &= \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \mathbf{B}_0^{-1})^{-1} \mathbf{X}^\top \mathbf{X} \\ &= \mathbf{B}_0^{-1} - \mathbf{B}_0^{-1}(\mathbf{X}^\top \mathbf{X} + \mathbf{B}_0^{-1})^{-1} \mathbf{B}_0^{-1} \\ &= \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \mathbf{B}_0^{-1})^{-1} \mathbf{B}_0^{-1}. \end{aligned}$$

Then,

$$\begin{aligned} \delta^* &= \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} \hat{\boldsymbol{\beta}} \\ &\quad - 2\hat{\boldsymbol{\beta}}^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} \boldsymbol{\beta}_0 \\ &= \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &\quad + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \end{aligned}$$

In a similar way for the second part,

$$\begin{aligned} (\mathbf{V}_0 + (\mathbf{X}^\top \mathbf{X})^{-1})^{-1} &= \mathbf{V}_0^{-1} - \mathbf{V}_0^{-1}(\mathbf{V}_0^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{V}_0^{-1} \\ &= \mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X}(\mathbf{V}_0^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \\ &= \mathbf{X}^\top \mathbf{X}((\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{V}_0)^{-1} \mathbf{V}_0^{-1}, \end{aligned}$$

we use these results and some algebra to show that $\mathbf{B}_0^\top \mathbf{V}_0^{-1} \mathbf{B}_0 + \hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X} \hat{\mathbf{B}} - \mathbf{B}_n^\top \mathbf{V}_n^{-1} \mathbf{B}_n = (\hat{\mathbf{B}} - \mathbf{B}_0)^\top \mathbf{V}_n (\hat{\mathbf{B}} - \mathbf{B}_0)$ taking into account that $\mathbf{V}_n = (\mathbf{V}_0^{-1} + \mathbf{X}^\top \mathbf{X})^{-1}$ and $\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$.

8. Show that in the linear regression model $\boldsymbol{\beta}_n^\top (\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1} \mathbf{M}^{-1} \mathbf{B}_n^{-1}) \boldsymbol{\beta}_n = \boldsymbol{\beta}_{**}^\top \mathbf{C} \boldsymbol{\beta}_{**}$ and $\boldsymbol{\beta}_{**} = \mathbf{X}_0 \boldsymbol{\beta}_n$.

Answer

Taking into account that $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} \mathbf{A}^{-1}$ [6], then we observe that $(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1} \mathbf{M}^{-1} \mathbf{B}_n^{-1}) = (\mathbf{B}_n + (\mathbf{X}_0^\top \mathbf{X}_0)^{-1})^{-1}$, where $(\mathbf{B}_n + (\mathbf{X}_0^\top \mathbf{X}_0)^{-1})^{-1} = \mathbf{X}_0^\top \mathbf{X}_0 - \mathbf{X}_0^\top \mathbf{X}_0 (\mathbf{B}_n^{-1} + \mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{X}_0 = \mathbf{X}_0^\top \mathbf{X}_0 - \mathbf{X}_0^\top \mathbf{X}_0 \mathbf{M}^{-1} \mathbf{X}_0^\top \mathbf{X}_0$, thus

$$\begin{aligned} \boldsymbol{\beta}_n^\top (\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1} \mathbf{M}^{-1} \mathbf{B}_n^{-1}) \boldsymbol{\beta}_n &= \boldsymbol{\beta}_n^\top (\mathbf{X}_0^\top \mathbf{X}_0 - \mathbf{X}_0^\top \mathbf{X}_0 \mathbf{M}^{-1} \mathbf{X}_0^\top \mathbf{X}_0) \boldsymbol{\beta}_n \\ &= \boldsymbol{\beta}_n^\top \mathbf{X}_0^\top (\mathbf{I}_{N_0} - \mathbf{X}_0 \mathbf{M}^{-1} \mathbf{X}_0^\top) \mathbf{X}_0 \boldsymbol{\beta}_n \\ &= \boldsymbol{\beta}_n^\top \mathbf{X}_0^\top \mathbf{C} \mathbf{X}_0 \boldsymbol{\beta}_n \\ &= \boldsymbol{\beta}_{**}^\top \mathbf{C} \boldsymbol{\beta}_{**}. \end{aligned}$$

Let's show that $\boldsymbol{\beta}_{**} = \mathbf{X}_0 \boldsymbol{\beta}_n$,

$$\begin{aligned}
\beta_{**} &= \mathbf{C}^{-1} \mathbf{X}_0 \mathbf{M}^{-1} \mathbf{B}_n^{-1} \beta_n \\
&= (\mathbf{I}_{N_0} + \mathbf{X}_0 \mathbf{B}_n \mathbf{X}_0^\top) \mathbf{X}_0 \mathbf{M}^{-1} \mathbf{B}_n^{-1} \beta_n \\
&= (\mathbf{I}_{N_0} + \mathbf{X}_0 \mathbf{B}_n \mathbf{X}_0^\top) \mathbf{X}_0 (\mathbf{B}_n - \mathbf{B}_n ((\mathbf{X}_0^\top \mathbf{X}_0)^{-1} + \mathbf{B}_n)^{-1} \mathbf{B}_n) \mathbf{B}_n^{-1} \beta_n \\
&= (\mathbf{I}_{N_0} + \mathbf{X}_0 \mathbf{B}_n \mathbf{X}_0^\top) (\mathbf{X}_0 \beta_n - \mathbf{X}_0 \mathbf{B}_n ((\mathbf{X}_0^\top \mathbf{X}_0)^{-1} + \mathbf{B}_n)^{-1} \beta_n) \\
&= \mathbf{X}_0 \beta_n - \mathbf{X}_0 \mathbf{B}_n ((\mathbf{X}_0^\top \mathbf{X}_0)^{-1} + \mathbf{B}_n)^{-1} \beta_n + \mathbf{X}_0 \mathbf{B}_n \mathbf{X}_0^\top \mathbf{X}_0 \beta_n \\
&\quad - \mathbf{X}_0 \mathbf{B}_n \mathbf{X}_0^\top \mathbf{X}_0 \mathbf{B}_n ((\mathbf{X}_0^\top \mathbf{X}_0)^{-1} + \mathbf{B}_n)^{-1} \beta_n \\
&= \mathbf{X}_0 \beta_n - \mathbf{X}_0 \mathbf{B}_n [((\mathbf{X}_0^\top \mathbf{X}_0)^{-1} + \mathbf{B}_n)^{-1} - \mathbf{X}_0^\top \mathbf{X}_0 + \mathbf{X}_0^\top \mathbf{X}_0 \mathbf{B}_n ((\mathbf{X}_0^\top \mathbf{X}_0)^{-1} + \mathbf{B}_n)^{-1}] \beta_n.
\end{aligned}$$

Using that $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{A} + \mathbf{B})^{-1}$, we observe that the expression in brackets is equal to $\mathbf{0}$, then we have the result.

9. Show that $(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \mathbf{S} + (\mathbf{B} - \hat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B} - \hat{\mathbf{B}})$ where $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$, $\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ in the multivariate regression model.

Answer

$$\begin{aligned}
(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) &= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} + \mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} + \mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B}) \\
&= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) + 2(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B}) \\
&\quad + (\mathbf{X}\mathbf{B} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{X}\mathbf{B} - \mathbf{X}\hat{\mathbf{B}}) \\
&= \mathbf{S} + (\mathbf{B} - \hat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B} - \hat{\mathbf{B}}),
\end{aligned}$$

given that $(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^\top (\mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B}) = \hat{\mathbf{U}}^\top \mathbf{X} (\hat{\mathbf{B}} - \mathbf{B})$, using that $\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ which implies $\mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}} = \mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}} + \mathbf{X}^\top \hat{\mathbf{U}}$, then $\mathbf{X}^\top \hat{\mathbf{U}} = \mathbf{0}$.

10. What is the probability that the Sun will rise tomorrow?

This is the most famous Richard Price's example developed in the Appendix of the Bayes' theorem paper [1]. Here, we implicitly use *Laplace's Rule of Succession* to solve this question. In particular, if we were a priori uncertain about the probability the Sun will rise on a specified day, we can assume a prior uniform distribution over $(0,1)$, that is, a beta $(1,1)$ distribution. Then, what is the probability that the Sun will rise tomorrow?

Answer

This exercise is an application of the Bernoulli-beta model. Thus, the likelihood is given by a binomial distribution where the probability of success is θ , $p(\mathbf{y}|\theta) \propto \theta^{\sum_{i=1}^N y_i} (1-\theta)^{N-\sum_{i=1}^N y_i}$. In addition, the prior distribution is beta, that is, $\pi(\theta) \propto \theta^{\alpha_0-1} (1-\theta)^{\beta_0-1}$, where $\alpha_0 = \beta_0 = 1$. Then, the predictive distribution that the sun will rise tomorrow is $p(Y_0 = 1|\mathbf{y}) = \frac{1+S}{2+N}$, where $S = \sum_{i=1}^N y_i$ is the number of successes (the Sun rise). $\frac{1+S}{2+N}$ is known

as the *Laplace's Rule of Succession* that was introduced by Laplace in the 18th century in the course of treating the sunrise problem.

11. Using information from Public Policy Polling in September 27th-28th for the 2016 presidential five-way race in USA, there are 411, 373 and 149 sampled people supporting Hillary Clinton, Donald Trump and other, respectively.
- Find the posterior probability of the percentage difference of people supporting Hillary versus Trump according to this data using a non-informative prior, that is, $\alpha_0 = [1 \ 1 \ 1]$ in the multinomial-Dirichlet model. What is the probability of having more supporters of Hillary vs Trump?
 - What is the probability that sampling one hundred independent individuals 44, 40 and 16 support Hillary, Trump and other, respectively?

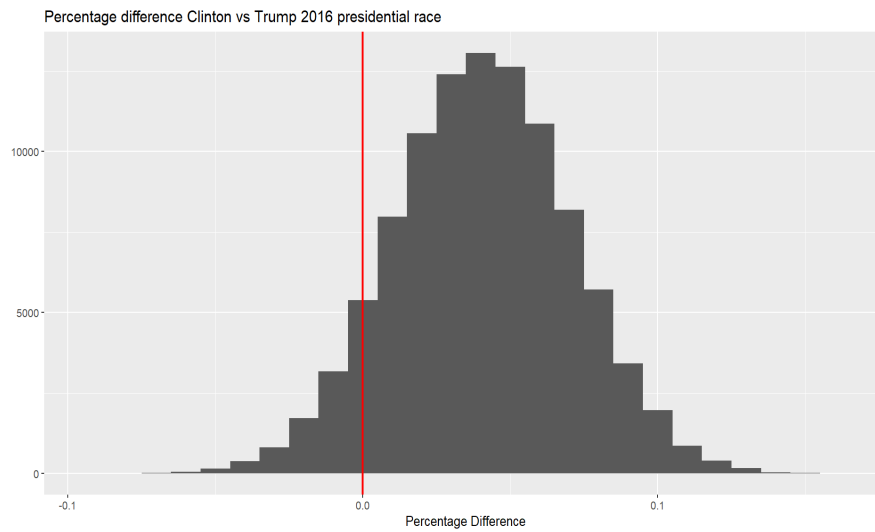
Answer

R code. Multinomial-Dirichlet model: Polling 2016 USA presidential race

```

1 set.seed(010101)
2 # Multinomial-Dirichlet example:
3 # Polling 2016 USA presidential race
4 y <- c(411, 373, 149)
5 # Clinton, Trump, Other
6 # Public Policy Polling September 27-28,
7 # 2016 five-way race
8 alpha0 <- rep(1, 3)
9 # Hyperparameters: non-informative distribution
10 alphan <- alpha0 + y
11 S <- 100000
12 # Sample draws of posterior
13 thetas <- MCMCpack::rdirichlet(S, alphan)
14 colnames(thetas) <- c("Clinton", "Trump", "Other")
15 head(thetas)
16      Clinton      Trump      Other
17 [1,] 0.4211346 0.4188607 0.1600046
18 [2,] 0.4244207 0.4224523 0.1531270
19 [3,] 0.4349268 0.3843953 0.1806779
20 [4,] 0.4533499 0.4005530 0.1460972
21 [5,] 0.4381799 0.3968502 0.1649699
22 [6,] 0.4436852 0.3971321 0.1591827
23 dif <- thetas[,1] - thetas[,2]
24 # Difference of shares Hillary vs Trump
25 data <- data.frame(dif)
26 names(data) <- c("Difference")
27 library(ggplot2)
28 p <- ggplot(data) +
29   geom_histogram(aes(x = Difference), binwidth = 0.01) +
30   geom_vline(xintercept=0.0, lwd=1, colour="red") +
31   ggtitle("Percentage difference Clinton vs Trump 2016
32     presidential race") + xlab("Percentage Difference") +
33     ylab("")
34 difmcmc <- coda::mcmc(dif)
35 # Declaring a MCMC object
36 summary(difmcmc)
37
38 Iterations = 1:1e+05
39 Thinning interval = 1
40 Number of chains = 1
41 Sample size per chain = 1e+05
42
43 1. Empirical mean and standard deviation for each
44 variable, plus standard error of the mean:
45
46      Mean          SD      Naive SE Time-series SE
47 4.062e-02    2.996e-02    9.474e-05    9.474e-05
48
49 2. Quantiles for each variable:
50
51      2.5%      25%      50%      75%      97.5%
52 -0.01817  0.02033  0.04058  0.06089  0.09923
53
54 CW <- mean(difmcmc>0)
55 CW
56 0.91339

```


**FIGURE 4.1**

Percentage difference: Hillary Clinton vs Donald Trump, five-way race.

There is a 95% probability that the percentage difference between Hillary and Trump according to this poll is (-1.8%, 9.9%). The probability of Hillary having more supporters is 91.3%

***R code. Multinomial-Dirichlet model: Polling 2016
USA presidential race***

```

1 # Predictive distribution by simulation
2 y0 <- c(44, 40, 16)
3 Pred <- apply(thetas, 1, function(p) {rmultinom(1, size =
4   sum(y0), prob = p)})
5 sum(sapply(1:S, function(s) {sum(Pred[,s] == y0) == 3}))/S
6 # Predictive distribution by analytical expression
7 PredY0 <- function(y0){
8   n <- sum(y0)
9   Res1 <- sum(sapply(1:length(y), function(l){lgamma(alphan[
10     l]+y0[l]) - lgamma(alphan[l])-lfactorial(y0[l]))}))
11   Res <- lfactorial(n)+lgamma(sum(alphan))-lgamma(sum(alphan
12     )+n) + Res1
13   return(exp(Res))
14 }
15 PredY0(y0)
16 0.00850

```

The probability that from one hundred random selected people 44 support Hillary, 40 support Trump and 16 support other candidate is 0.85%.

12. Math test example continues

You have a random sample of math scores of size $N = 50$ from a normal distribution, $Y_i \sim \mathcal{N}(\mu, \sigma)$. The sample mean and variance are equal to 102 and 10, respectively. Using the normal-normal/inverse-gamma model where $\mu_0 = 100$, $\beta_0 = 1$, $\alpha_0 = \delta_0 = 0.001$

- Get 95% confidence and credible intervals for μ .
- What is the posterior probability that $\mu > 103$?

Answer

R code. Math test example continues

```

1  set.seed(010101)
2  N <- 50
3  # Sample size
4  muhat <- 102
5  # Sample mean
6  sig2hat <- 10
7  # Sample variance
8  # Hyperparameters
9  mu0 <- 100
10 beta0 <- 1
11 delta0 <- 0.001
12 alpha0 <- 0.001
13 S <- 100000
14 # Posterior draws
15 alphan <- alpha0 + N
16 deltan <- sig2hat*(N - 1) + delta0 + beta0*N/(beta0 + N)*(
    muhat - mu0)^2
17 sig2Post <- invgamma::rinvgamma(S, shape = alphan, rate =
    deltan)
18 summary(sig2Post)
19 betan <- beta0 + N
20 mun <- (beta0*mu0 + N*muhat)/betan
21 muPost <- sapply(sig2Post, function(s2){rnorm(1, mun, sd = (
    s2/betan)^0.5)})
22 muPostq <- quantile(muPost, c(0.025, 0.5, 0.975))
23 muPostq
24      2.5%      50%      97.5%
25 101.0929 101.9625 102.8311
26 cutoff <- 103
27 PmuPostcutoff <- mean(muPost > cutoff)
28 PmuPostcutoff
29 0.00994
30 # Using Student's t
31 muPost_t <- ((deltan/(alphan*betan))^0.5)*rt(S, alphan) +
    mun
32 c1 <- rgb(173,216,230,max = 255, alpha = 50, names = "lt.
    blue")
33 c2 <- rgb(255,192,203, max = 255, alpha = 50, names = "lt.
    pink")
34 hist(muPost, main = "Histogram: Posterior mean", xlab = "
    Posterior mean", col = c2)
35 hist(muPost_t, main = "Histogram: Posterior mean", xlab = "
    Posterior mean", add = T, col = c1)
36 muPost_tq <- quantile(muPost_t, c(0.025, 0.5, 0.975))
37 muPost_tq
38      2.5%      50%      97.5%
39 101.0837 101.9608 102.8435
40 PmuPost_tcutoff <- mean(muPost_t > cutoff)
41 PmuPost_tcutoff
42 0.01087

```

We perform our calculations using the posterior conditional distribution, and the posterior marginal distribution. Both procedures give similar results as we can observe from Figure 4.2.

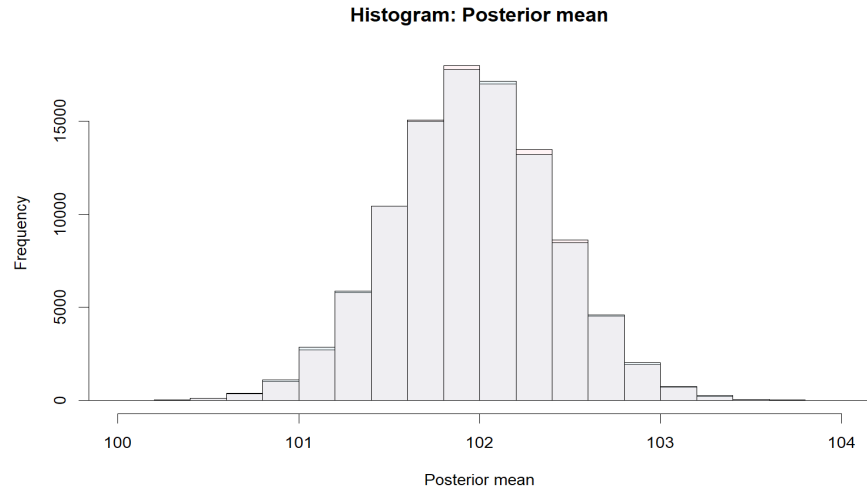


FIGURE 4.2

Histogram using the posterior conditional distribution and the posterior marginal distribution

We have that the 95% credible interval is (101.08, 102.84), and the probability of having a value greater than 103 is 1.09%.

13. Demand of electricity example continues

Set c_0 such that maximizes the marginal likelihood in the specifications with and without electricity price in the example of demand of electricity (empirical Bayes). Then, calculate the Bayes factor, and conclude if there is evidence supporting the inclusion of the price of electricity in the demand equation.

Answer

R code. Demand of electricity

```

1 rm(list = ls())
2 set.seed(010101)
3 # Electricity demand
4 DataUt <- read.csv("DataApplications/Utilities.csv", sep = "
5      ", header = TRUE, fileEncoding = "latin1")
6 DataUtEst <- DataUt %>%
7   filter(Electricity != 0)
8 attach(DataUtEst)
9 # Dependent variable: Monthly consumption (kWh) in log
10 Y <- log(Electricity)
11 N <- length(Y)
12 # Regressors quantity including intercept
13 X <- cbind(LnPriceElect, IndSocio1, IndSocio2, Altitude,
14           Nrooms, HouseholdMem, Children, Lnincome, 1)
15 # Regressor without price
16 Xnew <- cbind(IndSocio1, IndSocio2, Altitude, Nrooms,
17              HouseholdMem, Children, Lnincome, 1)
18 # Log marginal function (multiply by -1 due to minimization)
19 LogMarLikLM <- function(X, c0){
20   k <- dim(X)[2]
21   N <- dim(X)[1]
22   # Hyperparameters
23   B0 <- c0*diag(k)
24   b0 <- rep(0, k)
25   # Posterior parameters
26   bhat <- solve(t(X)%*%X)%*%t(X)%*%Y
27   # Force this matrix to be symmetric
28   Bn <- as.matrix(Matrix::forceSymmetric(solve(solve(B0) + t
29     (X)%*%X)))
30   bn <- Bn%*(solve(B0)%*%b0 + t(X)%*%X)%*%bhat
31   dn <- as.numeric(d0 + t(Y)%*%Y+t(b0)%*%solve(B0)%*%b0-t(bn
32     )%*%solve(Bn)%*%bn)
33   an <- a0 + N
34   # Log marginal likelihood
35   logpy <- (N/2)*log(1/pi)+(a0/2)*log(d0)-(an/2)*log(dn) +
36     0.5*log(det(Bn)/det(B0)) + lgamma(an/2)-lgamma(a0/2)
37   return(-logpy)
38 }
39 # Hyperparameters
40 d0 <- 0.001/2
41 a0 <- 0.001/2
42 # Empirical Bayes: Obtain c0 maximizing the log marginal
43   likelihood
44 c0 <- 1000
45 EB <- optim(c0, fn = LogMarLikLM, method = "Brent", lower =
46   0.0001, upper = 10^6, X = X)
47 EBnew <- optim(c0, fn = LogMarLikLM, method = "Brent", lower
48   = 0.0001, upper = 10^6, X = Xnew)
49 # Change of order to take into account the -1 in the
50   LogMarLikLM function
51 BFEM <- exp(EBnew$value - EB$value)
52 BFEM
53 71897938

```

The Bayes factor based on the empirical Bayes of the model with electricity price versus the model without electricity price is equal to 71897938, this gives very strong evidence to include the price in the specification.

14. Utility demand

Use the file *Utilities.csv* to estimate a multivariate linear regression model where $\mathbf{Y}_i = [\log(\text{electricity}_i) \log(\text{water}_i) \log(\text{gas}_i)]$ as function of $\log(\text{electricity price}_i)$, $\log(\text{water price}_i)$, $\log(\text{gas price}_i)$, IndSocio1_i , IndSocio2_i , Altitude_i , Nrooms_i , HouseholdMem_i , Children_i , and $\log(\text{Income}_i)$. Set a non-informative prior framework, $\mathbf{B}_0 = [0]_{11 \times 3}$, $\mathbf{V}_0 = 1000\mathbf{I}_{11}$, $\mathbf{\Psi}_0 = 1000\mathbf{I}_3$ and $\alpha_0 = 3$, where we have $K = 11$ (regressors plus intercept) and $M = 3$ (equations) in this exercise.

- (a) Find the posterior mean estimates and the highest posterior density intervals at 95% of \mathbf{B} and $\mathbf{\Sigma}$. Use the marginal distribution and the conditional distribution to obtain the posterior estimates of \mathbf{B} , and compare the results.
- (b) Find the Bayes factor comparing the baseline model in this exercise with the same specification but using the income in dollars. Now, calculate the Bayes factor using the income in thousand dollars. Is there any difference?
- (c) Find the predictive distribution for the monthly demand of electricity, water and gas in the baseline specification of a household located in the lowest socioeconomic condition in a municipality located below 1000 meters above the sea level, 2 rooms, 3 members with children, a monthly income equal to USD 500, an electricity price equal to USD/kWh 0.15, a water price equal to USD/M³ 0.70, and a gas price equal to USD/M³ 0.75.

Answer

We see that the posterior estimates of the location parameters based on the marginal distribution and the conditional distribution are very similar (conditional on $\mathbf{\Sigma}$). This is important as many times there is no analytical solutions in well-known forms of marginal posterior distributions, and consequently, we should get draws of the posterior distributions based on conditional distributions of block of parameters (See Chapter 5).

We find that the Bayes factor of the baseline model ($\log(\text{Income})$) versus the two alternative models using income in dollars and thousand dollars are 108925764 and 0.1089261. The former gives strong evidence in favor of the baseline model, whereas the latter gives positive evidence for the model using the income in thousand dollars. This result despite that the location coefficients are the same in the two alternative specifications, except for the change in scale of the coefficients associated with income. This example shows that Bayes factors are sensitive to units of measure, and

consequently, it is relevant to think carefully about the priors when performing hypothesis testing using a Bayesian framework. Observe that a nice feature in Bayesian inference is that we followed the same conceptual framework (Bayes factor) in the previous exercise and this exercise. In one hand, the previous exercise is an example of nested models, that is, one model is a restricted version of a more general model. On the other hand, this exercise is an example of non-nested models. This is not the case in the Frequentist approach. The statistical framework is not the same when testing nested and non-nested models.

*R code. Utilities demand: Multivariate regression,
posterior inference*

```

1 rm(list = ls())
2 set.seed(010101)
3 library(dplyr)
4 # Electricity demand
5 DataUt <- read.csv("DataApplications/Utilities.csv", sep = "
6      ", header = TRUE, fileEncoding = "latin1")
7 DataUtEst <- DataUt %>%
8   filter(Electricity != 0 & Water != 0 & Gas != 0)
9 attach(DataUtEst)
10 Y <- cbind(log(Electricity), log(Water), log(Gas))
11 X <- cbind(LnPriceElect, LnPriceWater, LnPriceGas, IndSocio1
12      , IndSocio2, Altitude, Nrooms, HouseholdMem, Children,
13      Lnincome, 1)
14 M <- dim(Y)[2]
15 K <- dim(X)[2]
16 N <- dim(Y)[1]
17 # Hyperparameters
18 B0 <- matrix(0, K, M)
19 c0 <- 1000
20 V0 <- c0*diag(K)
21 Psi0 <- c0*diag(M)
22 a0 <- M
23 # Posterior parameters
24 Bhat <- solve(t(X)%*%X)%*%t(X)%*%Y
25 S <- t(Y - X)%*%Bhat)%*%(Y - X)%*%Bhat)
26 Vn <- solve(solve(V0) + t(X)%*%X)
27 Bn <- Vn%(solve(V0)%*%B0 + t(X)%*%X)%*%Bhat)
28 Psiin <- Psi0 + S + t(B0)%*%solve(V0)%*%B0 + t(Bhat)%*%t(X)%*
29      %X)%*%Bhat - t(Bn)%*%solve(Vn)%*%Bn
30 an <- a0 + N
31 #Posterior draws
32 s <- 10000 #Number of posterior draws
33 SIGs <- replicate(s, LaplacesDemon::rinvwishart(an, Psiin))
34 BsCond <- sapply(1:s, function(s) {MixMatrix::rmatrixnorm(n
35      = 1, mean=Bn, U = Vn,V = SIGs[,s])})
36 summary(coda::mcmc(t(BsCond)))
37 Bs <- sapply(1:s, function(s) {MixMatrix::rmatrixt(n = 1,
38      mean=Bn, U = Vn,V = Psiin, df = an + 1 - M)})
39 summary(coda::mcmc(t(Bs)))
40 SIGMs <- t(sapply(1:s, function(l) {gdata::lowerTriangle(
41      SIGs[,l], diag=TRUE, byrow=FALSE)}))
42 summary(coda::mcmc(SIGMs))
43 hdiBs <- HDInterval::hdi(t(BsCond), credMass = 0.95) #
44      Highest posterior density credible interval
45 hdiBs
46 hdiSIG <- HDInterval::hdi(SIGMs, credMass = 0.95) # Highest
47      posterior density credible interval
48 hdiSIG
49
50

```


***R code. Utilities demand: Multivariate regression,
Bayes factors***

```

1 # Log marginal function (multiply by -1 due to minimization)
2 LogMarLikLM <- function(X, c0){
3   c10 <- c0[1]; c20 <- c0[2]
4   k <- dim(X)[2]
5   N <- dim(X)[1]
6   # Hyperparameters
7   V0 <- c10*diag(k)
8   Psi0 <- c20*diag(M)
9   # Posterior parameters
10  Bhat <- solve(t(X)%*%X)%*%t(X)%*%Y
11  S <- t(Y - X)%*%Bhat)%*%(Y - X)%*%Bhat)
12  Vn <- solve(solve(V0) + t(X)%*%X)
13  Bn <- Vn%*(solve(V0)%*%B0 + t(X)%*%X)%*%Bhat)
14  Psin <- Psi0 + S + t(B0)%*%solve(V0)%*%B0 + t(Bhat)%*%t(X)
    %*%X)%*%Bhat - t(Bn)%*%solve(Vn)%*%Bn
15  # Log marginal likelihood
16  logpy <- (N*M/2)*log(1/pi)+(a0/2)*log(det(Psi0)) - (an/2)*
    log(det(Psin)) + (M/2)*(log(det(Vn)) - log(det(V0))) +
    lgamma(an/2)-lgamma(a0/2)
17  return(-logpy)
18 }
19 c0 <- rep(1000, 2)
20 LogML <- LogMarLikLM(X=X, c0 = c0)
21 # Using income in dollars as regressor
22 Xnew <- cbind(LnPriceElect, LnPriceWater, LnPriceGas,
    IndSocio1, IndSocio2, Altitude, Nrooms, HouseholdMem,
    Children, exp(Lnincome), 1)
23 LogMLnew <- LogMarLikLM(X=Xnew, c0 = c0)
24 # Bayes factor
25 BF12 <- exp(LogMLnew - LogML)
26 BF12
27 # Using income in thousand dollars as regressor
28 XnewT <- cbind(LnPriceElect, LnPriceWater, LnPriceGas,
    IndSocio1, IndSocio2, Altitude, Nrooms, HouseholdMem,
    Children, exp(Lnincome)/1000, 1)
29 LogMLnewT <- LogMarLikLM(X=XnewT, c0 = c0)
30 # Bayes factor
31 BF13 <- exp(LogMLnewT - LogML)
32 BF13
33

```

R code. Utilities demand: Multivariate regression, predictive distribution

```

1 # Predictive distribution
2 Xpred <- c(log(0.15), log(0.70), log(0.75), 1, 0, 0, 2, 3,
3           1, log(500), 1)
4 Mean <- Xpred%%Bn
5 Hn <- 1+t(Xpred)%%Vn%%Xpred
6 UtilDemand <- exp(replicate(s, MixMatrix::rmatrixt(n = 1,
7             mean=Mean, U = Hn, V = Psin, df = an + 1 - M)))
8 ElePred <- UtilDemand[1,1,]
9 WatPred <- UtilDemand[1,2,]
10 GasPred <- UtilDemand[1,3,]
11 data <- data.frame(cbind(ElePred, WatPred, GasPred)) #Data
12           frame
13 annotations1 <- data.frame(
14   x = round(quantile(data$ElePred, c(0.025, 0.5, 0.975)),1),
15   y = c(600, 1000, 600),
16   label = c("2.5%", "50%", "97.5%:")
17 )
18 annotations2 <- data.frame(
19   x = round(quantile(data$WatPred, c(0.025, 0.5, 0.975)),1),
20   y = c(600, 1000, 600),
21   label = c("2.5%", "50%", "97.5%:")
22 )
23 annotations3 <- data.frame(
24   x = round(quantile(data$GasPred, c(0.025, 0.5, 0.975)),1),
25   y = c(600, 1000, 600),
26   label = c("2.5%", "50%", "97.5%:")
27 )
28 require(ggplot2) # Cool figures
29 require(ggpubr) # Multiple figures in one page
30 require(latex2exp) # LaTeX equations in figures
31 fig1 <- ggplot(data = data, aes(ElePred)) + geom_histogram(
32   bins = 40, color = "#000000", fill = "#0099F8") + xlab(
33     "kWh") + ylab("Frequency") + ggtitle("Electricity") +
34     xlim(0, 1050) + geom_text(data = annotations1, aes(x = x,
35       y = y, label = paste(label, x)), size = 3, fontface = "
36     bold")
37 fig2 <- ggplot(data = data, aes(WatPred)) + geom_histogram(
38   bins = 40, color = "#000000", fill = "#0099F8") + xlab(
39     TeX("$M^3$")) + ylab("Frequency") + ggtitle("Water") +
40     xlim(0, 100) + geom_text(data = annotations2, aes(x = x,
41       y = y, label = paste(label, x)), size = 3, fontface = "
42     bold")
43 fig3 <- ggplot(data = data, aes(GasPred)) + geom_histogram(
44   bins = 40, color = "#000000", fill = "#0099F8") + xlab(
45     TeX("$M^3$")) + ylab("Frequency") + ggtitle("Gas") +
46     xlim(0, 80) + geom_text(data = annotations3, aes(x = x,
47       y = y, label = paste(label, x)), size = 3, fontface = "
48     bold")

```

Figures 4.3, 4.4 and 4.5 show the marginal predictive distributions of electricity, water and gas for the reference household. The median predictive values are kWh 168.8, M³ 12.3 and M³ 10.1, respectively. In addition, the 95% credible intervals are (27.7, 1028.9), (1.5, 98.7) and (1.5, 67.5) for electricity, water and gas.

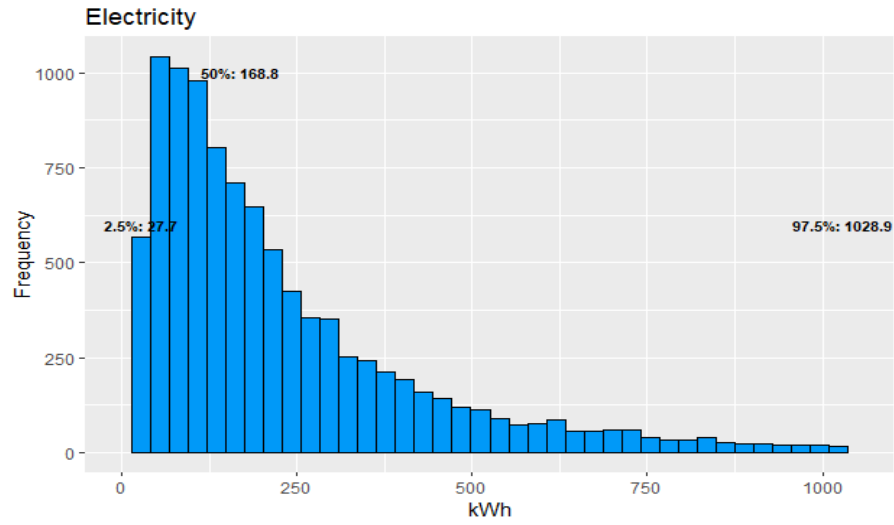
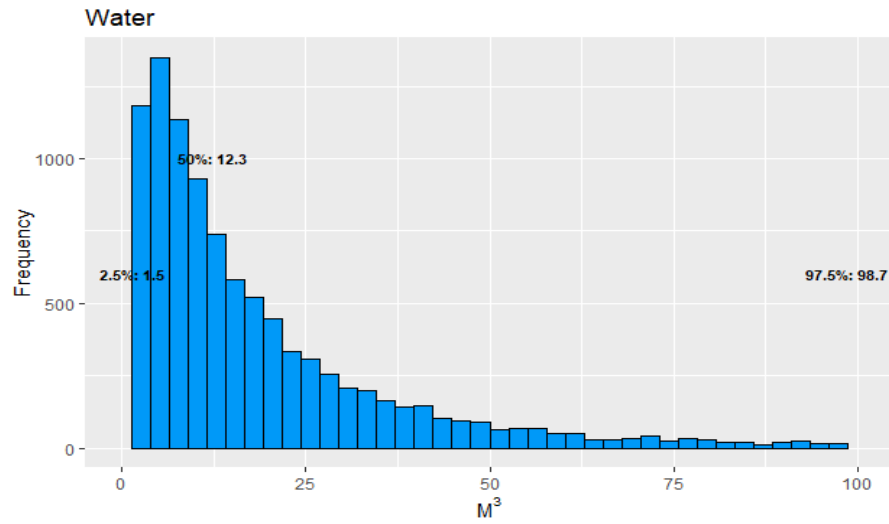
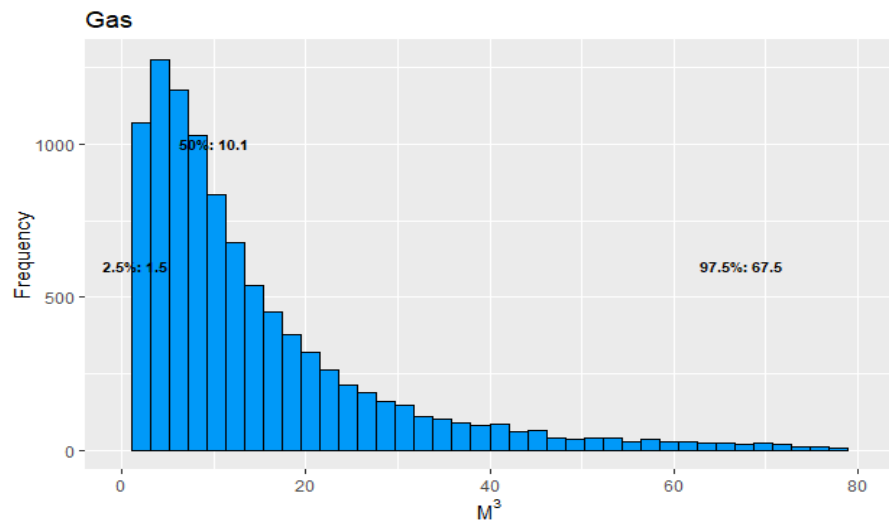


FIGURE 4.3

Histogram using the posterior predictive distribution of electricity demand

**FIGURE 4.4**

Histogram using the posterior predictive distribution of water demand

**FIGURE 4.5**

Histogram using the posterior predictive distribution of gas demand

5

Simulation methods

5.1 Solutions of Exercises

1. The inverse transform method ...

R code. Example: Math test

```
1 N <- 50 # Sample size
2 y_bar <- 102 # Sample mean
3 s2 <- 10 # Sample variance
4 alpha <- N - 1
5 serror <- (s2/N)^0.5
6 y.H0 <- c(100, 100.5, 101, 101.5, 102)
7 test <- (y.H0 - y_bar)/serror
8 pval <- 2*pt(test, alpha)
9 pval
10 0.0000459 0.0015431 0.0299338 0.2690040 1
11 # p-values
12 P001 <- (gamma(N/2)*((N-1)*serror^2)^(-0.5)*(1+test^2/alpha)
13 ^(-N/2))/(gamma(1/2)*gamma((N-1)/2))
14 P001/(1+P001)
15 0.0001705 0.0050345 0.0725330 0.3210223 0.4702050
16 # Posterior model probability of the null hypothesis.
```



Part II

Regression models: A GUIDed tour



6

Univariate models

6.1 Solutions of Exercises

1. Get the posterior conditional distributions of the Gaussian linear model assuming independent priors $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}) \times \pi(\sigma^2)$, where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}_0)$ and $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$.

Answer

The joint posterior distribution of the parameters is

$$\begin{aligned}
 \pi(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \mathbf{X}) \pi(\boldsymbol{\beta}) \pi(\sigma^2) \\
 &\propto (\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\
 &\quad \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \\
 &\quad \times \frac{(\delta_0/2)^{(\alpha_0/2)}}{\Gamma(\alpha_0/2)} \frac{1}{(\sigma^2)^{(\alpha_0/2+1)}} \exp \left\{ -\frac{\delta_0}{2\sigma^2} \right\} \\
 &\propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \\
 &\quad \times \frac{1}{(\sigma^2)^{(\alpha_0+N)/2+1}} \exp \left\{ -\frac{\delta_0 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right\} \\
 &= \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \times \underbrace{\frac{1}{(\sigma^2)^{(\frac{\alpha_n}{2}+1)}} \exp \left\{ -\frac{\delta_n}{2\sigma^2} \right\}}_1.
 \end{aligned}$$

Observe that (1) is the kernel of an inverse-gamma density function. Thus, $\sigma^2 | \boldsymbol{\beta}, \mathbf{y}, \mathbf{X} \sim IG(\alpha_n/2, \delta_n/2)$, where $\alpha_n = \alpha_0 + N$ and $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

Let's see the posterior distribution of β ,

$$\begin{aligned}
\pi(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto p(\mathbf{y} | \beta, \sigma^2, \mathbf{X}) \pi(\beta) \pi(\sigma^2) \\
&\propto (\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \right\} \\
&\times \exp \left\{ -\frac{1}{2} (\beta - \beta_0)^\top \mathbf{B}_0^{-1} (\beta - \beta_0) \right\} \\
&\times \frac{(\delta_0/2)^{(\alpha_0/2)}}{\Gamma(\alpha_0/2)} \frac{1}{(\sigma^2)^{(\alpha_0/2+1)}} \exp \left\{ -\frac{\delta_0}{2\sigma^2} \right\} \\
&= (\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{\sigma^{-2}}{2} [\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta] \right\} \\
&\times \exp \left\{ -\frac{1}{2} [\beta^\top \mathbf{B}_0^{-1} \beta - \beta^\top \mathbf{B}_0^{-1} \beta_0 - \beta_0^\top \mathbf{B}_0^{-1} \beta + \beta_0^\top \mathbf{B}_0^{-1} \beta_0] \right\} \\
&\times \frac{(\delta_0/2)^{(\alpha_0/2)}}{\Gamma(\alpha_0/2)} \frac{1}{(\sigma^2)^{(\alpha_0/2+1)}} \exp \left\{ -\frac{\delta_0}{2\sigma^2} \right\} \\
&\propto \exp \left\{ -\frac{1}{2} [\beta^\top (\mathbf{B}_0^{-1} + \sigma^{-2} \mathbf{X}^\top \mathbf{X}) \beta - 2\beta^\top (\mathbf{B}_0^{-1} \beta_0 + \sigma^{-2} \mathbf{X}^\top \mathbf{X} \hat{\beta})] \right\} \\
&\times \frac{1}{(\sigma^2)^{(\alpha_0+N)/2+1}} \exp \left\{ -\frac{\delta_0 + \mathbf{y}^\top \mathbf{y}}{2\sigma^2} \right\},
\end{aligned}$$

where $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

Adding and subtracting $\beta_n^\top \mathbf{B}_n^{-1} \beta_n$ where

$$\begin{aligned}
\mathbf{B}_n &= (\mathbf{B}_0^{-1} + \sigma^{-2} \mathbf{X}^\top \mathbf{X})^{-1} \\
\beta_n &= \mathbf{B}_n (\mathbf{B}_0^{-1} \beta_0 + \sigma^{-2} \mathbf{X}^\top \mathbf{X} \hat{\beta}) = \mathbf{B}_n (\mathbf{B}_0^{-1} \beta_0 + \sigma^{-2} \mathbf{X}^\top \mathbf{y}),
\end{aligned}$$

and completing the square

$$\begin{aligned}
\pi(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto \exp \left\{ -\frac{1}{2} [\beta^\top (\mathbf{B}_0^{-1} + \sigma^{-2} \mathbf{X}^\top \mathbf{X}) \beta - 2\beta^\top \mathbf{B}_n^{-1} \mathbf{B}_n (\mathbf{B}_0^{-1} \beta_0 + \sigma^{-2} \mathbf{X}^\top \mathbf{X} \hat{\beta}) \right. \\
&\quad \left. + \beta_n^\top \mathbf{B}_n^{-1} \beta_n - \beta_n^\top \mathbf{B}_n^{-1} \beta_n] \right\} \times \frac{1}{(\sigma^2)^{(\alpha_0+N)/2+1}} \exp \left\{ -\frac{\delta_0 + \mathbf{y}^\top \mathbf{y}}{2\sigma^2} \right\} \\
&= \exp \left\{ -\frac{1}{2} [\beta^\top \mathbf{B}_n^{-1} \beta - 2\beta^\top \mathbf{B}_n^{-1} \beta_n + \beta_n^\top \mathbf{B}_n^{-1} \beta_n] \right\} \\
&\times \frac{1}{(\sigma^2)^{(\alpha_0+N)/2+1}} \exp \left\{ -\frac{\delta_0 + \mathbf{y}^\top \mathbf{y} - \sigma^2 \beta_n^\top \mathbf{B}_n^{-1} \beta_n}{2\sigma^2} \right\} \\
&= \exp \left\{ -\frac{1}{2} (\beta - \beta_n)^\top \mathbf{B}_n^{-1} (\beta - \beta_n) \right\} \\
&\quad \underbrace{\hspace{10em}}_1 \\
&\times (\sigma^2)^{-(\frac{\alpha_n}{2}+1)} \exp \left\{ -\frac{\delta^*}{2\sigma^2} \right\},
\end{aligned}$$

where $\delta^* = \delta_0 + \mathbf{y}^\top \mathbf{y} + \sigma^2 \boldsymbol{\beta}_n^\top \mathbf{B}_n^{-1} \boldsymbol{\beta}_n$ does not depend on $\boldsymbol{\beta}$.

We can see that (1) is the kernel of a multivariate normal distribution with mean equal to $\boldsymbol{\beta}_n$ and covariance matrix \mathbf{B}_n , that is, $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim N(\boldsymbol{\beta}_n, \mathbf{B}_n)$.

We see that the posterior distributions are from the same family as the prior distributions.

2. Show that the posterior conditional distributions of the Gaussian linear model with heteroskedasticity assuming independent priors $\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}) = \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \times \prod_{i=1}^N \pi(\tau_i)$, where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}_0)$, $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$ and $\tau_i \sim G(v/2, v/2)$ are $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X} \sim N(\boldsymbol{\beta}_n, \mathbf{B}_n)$, $\sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X} \sim IG(\alpha_n, \delta_n)$ and $\tau_i|\boldsymbol{\beta}, \sigma^2, \mathbf{y}, \mathbf{X} \sim G(v_{1n}, v_{2in})$, where $\boldsymbol{\tau} = [\tau_1, \dots, \tau_n]^\top$, $\mathbf{B}_n = (\mathbf{B}_0^{-1} + \sigma^{-2} \mathbf{X}^\top \Psi \mathbf{X})^{-1}$, $\boldsymbol{\beta}_n = \mathbf{B}_n(\mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \sigma^{-2} \mathbf{X}^\top \Psi \mathbf{y})$, $\alpha_n = \alpha_0 + N$, $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \Psi (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, $v_{1n} = v + 1$, $v_{2in} = v + \sigma^{-2}(\mathbf{y}_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$, and $\Psi = \text{diagonal}\{\tau_i\}$.

Answer

The joint posterior distribution of the parameters is

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}|\mathbf{y}, \mathbf{X}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}, \mathbf{X}) \pi(\boldsymbol{\beta}) \pi(\sigma^2) \prod_{i=1}^N \pi(\tau_i) \\ &\propto \left(\prod_{i=1}^N \tau_i^{1/2} \right) (\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \Psi (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ &\quad \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \\ &\quad \times \frac{1}{(\sigma^2)^{(\alpha_0/2+1)}} \exp \left\{ -\frac{\delta_0}{2\sigma^2} \right\} \prod_{i=1}^N \tau_i^{v/2-1} \exp \{-v\tau_i/2\}. \end{aligned}$$

Thus, the conditional posterior distribution of $\sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X}$ is given by

$$\pi(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X}) \propto (\sigma^2)^{-\frac{N+\alpha_0}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \Psi (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \delta_0 \right\}.$$

This is the kernel of an inverse-gamma distribution with shape parameter α_n and rate parameter δ_n .

The conditional posterior distribution of $\tau_i|\boldsymbol{\beta}, \sigma^2, \mathbf{y}, \mathbf{X}$ is

$$\pi(\tau_i|\boldsymbol{\beta}, \sigma^2, \mathbf{y}, \mathbf{X}) \propto \tau_i^{(v+1)/2-1} \exp \left\{ -\frac{\tau_i}{2} [\sigma^{-2}(\mathbf{y}_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + v] \right\}.$$

The conditional posterior distribution of $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X}$ is given by

$$\pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X}) \propto \exp \left\{ -\frac{1}{2} [\sigma^{-2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \Psi(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{B}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \right\}.$$

Following same steps as in the previous exercise we get $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}, \mathbf{y}, \mathbf{X} \sim N(\boldsymbol{\beta}_n, \mathbf{B}_n)$.

3. The market value of soccer players in Europe continues

Use the setting of the previous exercise to perform inference using a Gibbs sampling algorithm of the the market value of soccer players in Europe setting $v = 5$ and same other hyperparameters as the homoscedastic case. Is there any meaningful difference for the coefficient associated with the national team compared to the application in the homoscedastic case?

Answer

R. code. The value of soccer players, programming our Gibbs sampler (heteroskedastic case)

```

1 rm(list = ls())
2 set.seed(010101)
3 ##### Linear regression: Value of
   soccer players #####
4 Data <- read.csv("DataApplications/1ValueFootballPlayers.csv",
   sep = ",", header = TRUE, fileEncoding = "latin1")
5 attach(Data)
6 y <- log(Value)
7 # Value: Market value in Euros (2017) of soccer players
8 # Regressors quantity including intercept
9 X <- cbind(1, Perf, Age, Age2, NatTeam, Goals, Exp, Exp2)
10 # Perf: Performance. Perf2: Performance squared. Age: Age;
   Age: Age squared.
11 # NatTeam: Indicator of national team. Goals: Scored goals.
   Goals2: Scored goals squared
12 # Exp: Years of experience. Exp2: Years of experience
   squared. Assists: Number of assists
13 k <- dim(X)[2]
14 N <- dim(X)[1]
15 # Hyperparameters
16 d0 <- 0.001/2
17 a0 <- 0.001/2
18 b0 <- rep(0, k)
19 c0 <- 1000
20 B0 <- c0*diag(k)
21 B0i <- solve(B0)
22 v <- 5
23 # MCMC parameters
24 mcmc <- 5000
25 burnin <- 5000
26 tot <- mcmc + burnin
27 thin <- 1
28 # Posterior distributions programming the Gibbs sampling
29 # Auxiliary parameters
30 an <- a0 + N
31 v1n <- v + 1
32 # Gibbs sampling functions
33 PostSig2 <- function(Beta, tau){
34   dn <- d0 + t(y - X%%Beta)%%diag(tau)%(y - X%%Beta)
35   sig2 <- invgamma::rinvgamma(1, shape = an/2, rate = dn/2)
36   return(sig2)
37 }
38 PostBeta <- function(sig2, tau){
39   Bn <- solve(B0i + sig2^(-1)*t(X)%%diag(tau)%%X)
40   bn <- Bn%%(B0i%%b0 + sig2^(-1)*t(X)%%diag(tau)%%y)
41   Beta <- MASS::mvrnorm(1, bn, Bn)
42   return(Beta)
43 }
44 PostTau <- function(sig2, Beta, i){
45   v2n <- v + sig2^(-1)*(y[i]-X[i,]%%Beta)^2
46   tau_i <- rgamma(1, v1n, v2n)
47   return(tau_i)
48 }

```

R. code. The value of soccer players, programming our Gibbs sampler (heteroskedastic case)

```

1 PostBetas <- matrix(0, mcmc+burnin, k)
2 PostSigma2 <- rep(0, mcmc+burnin)
3 Beta <- rep(0, k)
4 tau <- rep(1, N)
5 for(s in 1:tot){
6   sig2 <- PostSig2(Beta = Beta, tau = tau)
7   PostSigma2[s] <- sig2
8   Beta <- PostBeta(sig2 = sig2, tau = tau)
9   PostBetas[s,] <- Beta
10  tau <- sapply(1:N, function(i){PostTau(sig2 = sig2, Beta =
    Beta, i)})
11 }
12 keep <- seq((burnin+1), tot, thin)
13 PosteriorBetas <- PostBetas[keep,]
14 colnames(PosteriorBetas) <- c("Intercept", "Perf", "Age", "
    Age2", "NatTeam", "Goals", "Exp", "Exp2")
15 summary(coda::mcmc(PosteriorBetas))
16 PosteriorSigma2 <- PostSigma2[keep]
17 summary(coda::mcmc(PosteriorSigma2))
18 summary(coda::mcmc(exp(PosteriorBetas[,5])-1))
19 Iterations = 1:5000
20 Thinning interval = 1
21 Number of chains = 1
22 Sample size per chain = 5000
23 1. Empirical mean and standard deviation for each variable,
24 plus standard error of the mean:
25 Mean          SD          Naive SE Time-series SE
26 1.246823      0.235881      0.003336      0.004010
27 2. Quantiles for each variable:
28 2.5%   25%   50%   75% 97.5%
29 0.815 1.086 1.238 1.395 1.744

```

We see in this application that the value of a top soccer player in Europe increases 124% ($\exp(0.80) - 1$) on average when he has played in the national team, the credible interval at 95% is (81%, 174%). These values are not very different from the application assuming homoscedasticity in the book.

4. Example: Determinants of hospitalization continues

Program a Gibbs sampling algorithm in the application of determinants of hospitalization.

Answer

*R. code. Determinants of hospitalization,
programming our Gibbs sampler*

```

1 set.seed(010101)
2 Data <- read.csv("DataApplications/2HealthMed.csv", sep = ",
   ", header = TRUE, fileEncoding = "latin1")
3 attach(Data)
4 str(Data)
5 y <- Hosp # Dependent variables
6 X <- cbind(1, SHI, Female, Age, Age2, Est2, Est3, Fair, Good
   , Excellent) # Regressors
7 K <- dim(X)[2]
8 N <- dim(X)[1]
9 # Hyperparameters
10 b0 <- rep(0, K) # Prio mean
11 B0 <- diag(K) # Prior covariance
12 B0i <- solve(B0)
13 mcmc <- 1000; burnin <- 500; thin <- 2; tot <- mcmc + burnin
   ; keep <- seq(burnin, tot, thin)
14 # Posterior distributions programming the Gibbs sampling
15 # Auxiliary parameters
16 XtX <- t(X)%*%X
17 # Gibbs sampling functions
18 PostBeta <- function(Yl){
19   Bn <- solve(B0i + XtX)
20   bn <- Bn%*%(B0i%*%b0 + t(X)%*%Yl)
21   Beta <- MASS::mvrnorm(1, bn, Bn)
22   return(Beta)
23 }
24 PostYl <- function(Beta, i){
25   Ylmean <- X[i,]%*%Beta
26   if(y[i] == 1){
27     Yli <- truncnorm::rtruncnorm(1, a = 0, b = Inf, mean =
       Ylmean, sd = 1)
28   }else{
29     Yli <- truncnorm::rtruncnorm(1, a = -Inf, b = 0, mean =
       Ylmean, sd = 1)
30   }
31   return(Yli)
32 }
33 PostBetas <- matrix(0, mcmc+burnin, K)
34 Beta <- rep(0, K)
35 # create progress bar in case that you want to see
   iterations progress
36 pb <- winProgressBar(title = "progress bar", min = 0, max =
   tot, width = 300)
37 for(s in 1:tot){
38   Yl <- sapply(1:N, function(i){PostYl(Beta = Beta, i)})
39   Beta <- PostBeta(Yl = Yl)
40   PostBetas[s,] <- Beta
41   setWinProgressBar(pb, s, title=paste( round(s/tot*100, 0),
     "% done"))
42 }
43 close(pb)
44 keep <- seq((burnin+1), tot, thin)
45 PosteriorBetas <- PostBetas[keep,]
46 colnames(PosteriorBetas) <- c("Intercept", "SHI", "Female",
   "Age", "Age2", "Est2", "Est3", "Fair", "Good", "
   Excellent")
47 summary(coda::mcmc(PosteriorBetas))

```



Part III

Advanced methods: Theory, applications and programming



Bibliography

- [1] Thomas Bayes. LII. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, 53:370–418, 1763.
- [2] H. Jeffreys. *Theory of Probability*. Oxford University Press, London, 1961.
- [3] Valen E Johnson and David Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(2):143–170, 2010.
- [4] Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.
- [5] Kevin P Murphy. Conjugate bayesian analysis of the Gaussian distribution. *def*, 1(2 σ 2):16, 2007.
- [6] A. F. M. Smith. A General Bayesian Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological)*., 35(1):67–75, 1973.
- [7] A. Zellner. *Introduction to Bayesian inference in econometrics*. John Wiley & Sons Inc., 1996.