Half Title

Title Page

LOC Page



Contents

Fc	preword	ix
Pı	reface	хi
C	ontributors	xiii
$\mathbf{S}_{\mathbf{J}}$	ymbols	$\mathbf{x}\mathbf{v}$
Ι	Foundations: Theory, simulation methods and programming	1
1	Solutions of chapter 1 Basic formal concepts 1.1 Solutions Exercises	3 3
2	Solutions of chapter 2 Conceptual differences: Bayesian and Frequentist approaches 2.1 Solutions Exercises	23 23
3	Solutions of chapter 2 Cornerstone models: Conjugate families 3.1 Solutions Exercises	27 27
Bi	ibliography	37

Foreword

Preface

Contributors

Symbols

Symbol Description

\neg	Negation symbol.	${\cal R}$	The Real set.
\propto	Proportional symbol.	Ø	Empty set.
L	Independence symbol.	1	Indicator function

Part I

Foundations: Theory, simulation methods and programming

Solutions of chapter 1 Basic formal concepts

1.1 Solutions 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:

$$P(B|WB) = \frac{P(B,WB)}{P(WB)}$$

$$= \frac{P(WB|B) \times P(B)}{P(WB|B) \times P(B) + (1 - P(WB|B)) \times (1 - P(B))}$$

$$= \frac{0.8 \times 0.15}{0.8 \times 0.15 + 0.2 \times 0.85}$$

$$= 0.41$$
(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,

$$P(C_{i}|H_{3}, P_{1}) = \frac{P(C_{i}, H_{3}, P_{1})}{P(H_{3}, P_{1})}$$

$$= \frac{P(H_{3}|C_{i}, P_{1})P(C_{i}|P_{1})P(P_{1})}{P(H_{3}|P_{1}) \times P(P_{1})}$$

$$= \frac{P(H_{3}|C_{i}, P_{1})P(C_{i})}{P(H_{3}|P_{1})}$$

$$= \frac{1/2 \times 1/4}{1/3}$$

$$= \frac{3}{8},$$

$$(1.2)$$

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:

$$P\left(C_{n}|(H_{2}\cup\ldots\cup H_{n-1})\cap P_{1}\right) = \frac{P\left((H_{2}\cup H_{3}\cup\ldots\cup H_{n-1})|C_{n}\cap P_{1}\right)P(C_{n}|P_{1})P(P_{1})}{P\left((H_{2}\cup H_{3}\cup\ldots\cup H_{n-1})|P_{1}\right)P(P_{1})}$$

$$= \frac{\left[P\left(H_{2}|C_{n}\cap P_{1}\right)+\ldots P\left(H_{n-1}|C_{n}\cap P_{1}\right)\right]P(C_{n})}{P\left(H_{2}|P_{1}\right)+P\left(H_{3}|P_{1}\right)+\ldots+P\left(H_{n-1}|P_{1}\right)}$$

$$= \frac{1\times\left(\frac{1}{n}\right)}{\frac{1}{n-1}+\frac{1}{n-1}+\ldots+\frac{1}{n-1}}$$

$$= \left(\frac{1}{n}\right)\left(\frac{n-1}{n-2}\right).$$
(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

```
\operatorname{set.seed}(0101) # Set simulation seed
S < - \ 100000 \ \# \ \mathit{Simulations}
 \begin{array}{lll} \text{Game} & \leftarrow & \text{function} \, (\, \text{opt} \, = \, 3\,) \, \{ \\ \# & \textit{opt: number of options. opt} \, > \, 2 \end{array} 
opts <- 1:opt
car \leftarrow sample(opts, 1) \# car location
guess1 <- sample(opts, 1) # Initial guess
if(opt == 3 \&\& car != guess1) {
 host \leftarrow opts[-c(car, guess1)]
 } else {
 host \leftarrow sample(opts[-c(car, guess1)], 1)
win1 <- guess1 == car # Win given no change
if(opt == 3) {
 guess2 <- opts[-c(host, guess1)]</pre>
 } else {
 guess2 \leftarrow sample(opts[-c(host, guess1)], 1)
win2 <- guess2 == car # Win given change
return (c(win1, win2))
#Win probabilities
Prob <- rowMeans(replicate(S, Game(opt = 4)))
#Winning probabilities no changing door
Prob[1]
0.25151
#Winning probabilities changing door
Prob [2]
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,

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

$$= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} e^{-\lambda \beta_0}\right) \left(\frac{\lambda^{\sum_{i=1}^N y_i} e^{-N\lambda}}{\prod_{i=1}^N y_i!}\right)$$

$$= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{1}{\prod_{i=1}^N y_i!}\right) \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} e^{-\lambda(\beta_0 + N)}$$

$$\propto \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} e^{-\lambda(\beta_0 + N)}. \tag{1.5}$$

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

$$E[\lambda|\mathbf{y}] = \frac{\alpha_n}{\beta_n}$$

$$= \frac{\sum_{i=1}^{N} y_i + \alpha_0}{\beta_0 + N}$$

$$= \left(\frac{N}{\beta_0 + N}\right) \bar{y} + \left(\frac{\beta_0}{\beta_0 + N}\right) \frac{\alpha_0}{\beta_0}$$

$$= w\bar{y} + (1 - w) E[\lambda],$$

$$(1.6)$$

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

$$\pi (Y_0|\mathbf{y}) = \int_0^\infty \frac{\lambda^{y_0} e^{-\lambda}}{y_0!} \pi (\lambda|\mathbf{y}) d\lambda$$

$$= \int_0^\infty \left(\frac{\lambda^{y_0} e^{-\lambda}}{y_0!}\right) \left(\frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \lambda^{\alpha_n - 1} e^{-\lambda \beta_n}\right) d\lambda$$

$$= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n) y_0!} \int_0^\infty \lambda^{y_0 + \alpha_n - 1} e^{-\lambda (1 + \beta_n)} d\lambda$$

$$= \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n) y_0!} \frac{\Gamma(y_0 + \alpha_n)}{(1 + \beta_n)^{y_0 + \alpha_n}}$$

$$= \frac{\Gamma(y_0 + \alpha_n)}{\Gamma(\alpha_n) y_0!} \left(\frac{1}{1 + \beta_n}\right)^{y_0} \left(\frac{\beta_n}{1 + \beta_n}\right)^{\alpha_n}$$

7

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

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

$$\left[\hat{\alpha_0}\hat{\beta_0}\right] = \arg\max_{\alpha_0,\beta_0} \ln(p(\mathbf{y})),$$

where

$$p(y) = \int_0^\infty \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} e^{-\lambda \beta_0}\right) \left(\prod_{i=1}^N \frac{\lambda^{y_0} e^{-\lambda}}{y_0!}\right) d\lambda$$

$$= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0) \prod_{i=1}^N y_i!}\right) \int_0^\infty \lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1} e^{-\lambda(\beta_0 + N)} d\lambda$$

$$= \left(\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0) \prod_{i=1}^N y_i!}\right) \left(\frac{\Gamma\left(\sum_{i=1}^N y_i + \alpha_0\right)}{(\beta_0 + N)^{\sum_{i=1}^N y_i + \alpha_0}}\right)$$

$$= \frac{\Gamma\left(\sum_{i=1}^N y_i + \alpha_0\right)}{\Gamma(\alpha_0) \prod_{i=1}^N y_i!} \left(\frac{1}{\beta_0 + N}\right)^{\sum_{i=1}^N y_i} \left(\frac{\beta_0}{\beta_0 + N}\right)^{\alpha_0}.$$
(1.8)

R code. Health insurance, predictive distribution using vague hyperparameters

```
set.seed(010101)
y < -c(0, 3, 2, 1, 0) \# Data
N \leftarrow length(y)
# Predictive distribution
ProbBo <- function(y, a0, b0){
         N \leftarrow length(y)
         \#sample \ size
         aN \leftarrow a0 + sum(y)
         # Posterior shape parameter
         bN \leftarrow b0 + N
         \# Posterior scale parameter
         p < -1 / (bN + 1)
         # Probability negative binomial density
         Pr \leftarrow 1 - pnbinom(0, size = aN, prob = (1 - p))
         # Probability of visiting the Doctor
# Observe that in R there is a slightly
         \# \ different \ parametrization .
         return (Pr)
}
# Using a vague prior:
PriMeanV <- a0 / b0 # Prior mean
PriVarV <- a0 / b0^2 # Prior variance
\mathbf{P}\mathbf{p}
0.67
```

R code. Health insurance, predictive distribution using empirical Bayes

```
\# Using Emprirical Bayes
LogMgLik <- \ function (theta , \ y) \{
        N \leftarrow length(y)
         \#sample \ size
         a0 <- theta[1]
         # prior shape hyperparameter
         b0 < - theta[2]
        # prior scale hyperparameter
aN <- sum(y) + a0</pre>
         # posterior shape parameter
         if(a0 \le 0 \mid \mid b0 \le 0)
                  \#A\,voiding\ negative\ values
                  lnp < -Inf
         else \{lnp \leftarrow lgamma(aN) - sum(y)*log(b0+N) +
         a0*log(b0/(b0+N)) - lgamma(a0)
    # log marginal likelihood
         return(-lnp)
}
theta0 < c(0.01, 0.01)
\# Initial values
control <- list (maxit = 1000)
\# Number of iterations in optimization
EmpBay <- optim(theta0, LogMgLik, method = "BFGS",
control = control, hessian = TRUE, y = y)
# Optimization
EmpBay$convergence
\# Checking convergence
EmpBay$value # Maximum
a0EB <- EmpBay$par[1]
# Prior shape using empirical Bayes
a0EB
128.383
b0EB <- EmpBay$par[2]
# Prior scale using empirical Bayes
b0EB
106.801
PriMeanEB \leftarrow a0EB / b0EB
# Prior mean
PriVarEB <- a0EB / b0EB^2
# Prior variance
\stackrel{\sim}{PpEB} \leftarrow ProbBo(y, a0 = a0EB, b0 = b0EB)
# This setting is using emprical Bayes.
PpEB
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, $Y_i \sim Ber(p)$, 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

$$\pi(p|\mathbf{y}) = \left[\frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p^{\alpha_0 - 1} (1 - p)^{\beta_0 - 1}\right] \left[\prod_{i=1}^N p^{y_i} (1 - p)^{1 - y_i}\right]$$

$$= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p^{\sum_{i=1}^N y_i + \alpha_0 - 1} (1 - p)^{\beta_0 + N - \sum_{i=1}^N y_i - 1}$$

$$\propto p^{\sum_{i=1}^N y_i + \alpha_0 - 1} (1 - p)^{\beta_0 + N - \sum_{i=1}^N y_i - 1}.$$
(1.9)

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

$$E[p|\mathbf{y}] = \frac{\alpha_n}{\alpha_n + \beta_n}$$

$$= \frac{\sum_{i=1}^{N} y_i + \alpha_0}{\alpha_0 + \beta_0 + N}$$

$$= \frac{N\bar{y}}{\alpha_0 + \beta_0 + N} + \frac{\alpha_0}{\alpha_0 + \beta_0 + N}$$

$$= \frac{N}{\alpha_0 + \beta_0 + N} (\bar{y}) + \frac{\alpha_0 + \beta_0}{\alpha_0 + \beta_0 + N} \left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right)$$

$$= w(\bar{y}) + (1 - w)E[p],$$
(1.10)

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

$$\pi(Y_0 = 1|\mathbf{y}) = \int_0^1 P(Y_0 = 1|\mathbf{y}, p)\pi(p|\mathbf{y}) dp$$

$$= \int_0^1 p \times \pi(p|\mathbf{y}) dp$$

$$= \mathbb{E}[p|\mathbf{y}]$$

$$= \frac{\alpha_n}{\alpha_n + \beta_n}.$$
(1.11)

To use empirical Bayes, we have the following setting

$$\left[\hat{\alpha_0}\hat{\beta_0}\right] = \arg\max_{\alpha_0,\beta_0} \ln(p(\mathbf{y})),$$

where

$$p(\mathbf{y}) = \int_{0}^{1} \left[\frac{\Gamma(\alpha_{0} + \beta_{0})}{\Gamma(\alpha_{0}) \Gamma(\beta_{0})} p^{\alpha_{0} - 1} (1 - p)^{\beta_{0} - 1} \right] \left[\prod_{i=1}^{N} (1 - p)^{1 - y_{i}} \right] dp \quad (1.12)$$

$$= \frac{\Gamma(\alpha_{0} + \beta_{0})}{\Gamma(\alpha_{0}) \Gamma(\beta_{0})} \int_{0}^{1} p^{\sum_{i=1}^{N} y_{i} + \alpha_{0} - 1} (1 - p)^{\beta_{0} + N - \sum_{i=1}^{N} y_{i} - 1} dp$$

$$= \frac{\Gamma(\alpha_{0} + \beta_{0})}{\Gamma(\alpha_{0}) \Gamma(\beta_{0})} \frac{\Gamma(\sum_{i=1}^{N} y_{i} + \alpha_{0}) \Gamma(\beta_{0} + N - \sum_{i=1}^{N} y_{i})}{\Gamma(\alpha_{0} + \beta_{0} + N)}.$$

$R\ code.\ Car\ claim,\ predictive\ distribution\ using\ vague\ hyperparameters$

```
set.seed(010101)
y \ \leftarrow \ c \ (0 \ , \ 1 \ , \ 0 \ , \ 1 \ , \ 1 \ , \ 0 \ , \ 1 \ , \ 0 \ , \ 1 \ , \ 1 \ , \ 0)
# Data
N \leftarrow length(y)
#require (TailRank)
# Predictive distribution
ProbBo <- function (y, a0, b0) {
            N \leftarrow length(y)
            #sample size
            aN \leftarrow a0 + sum(y)
            # Posterior shape parameter bN <- b0 + N - sum(y)
             # Posterior scale parameter
             \begin{array}{l} \overset{''}{\operatorname{pr}} < - \ aN \ / \ (aN \ + \ bN) \\ \# \ \mathit{Probability} \ \mathit{of} \ \mathit{a} \ \mathit{claim} \ \mathit{the} \ \mathit{next} \ \mathit{year} \end{array} 
             return (pr)
}
# Using a vague prior:
a0 <- 1 # Prior shape parameter
b0 \leftarrow 1 \# Prior scale parameter
PriMeanV \leftarrow a0 / (a0 + b0)
# Prior mean
PriVarV \leftarrow (a0*b0) / (((a0+b0)^2)*(a0+b0+1))
# Prior variance
Pp \leftarrow ProbBo(y, a0 = 1, b0 = 1)
# This setting is defining vague prior information.
\# The probability of a claim
0.47
```

R code. Car claim, predictive distribution using empirical Bayes

```
# Using Emprirical Bayes
LogMgLik <- function(theta, y){
                        N < - length(y)
                        #sample size
                        a0 <- theta[1]
                         # prior shape hyperparameter
                         b0 \leftarrow theta[2]
                         # prior scale hyperparameter
                        aN \leftarrow sum(y) + a0
                        \# posterior shape parameter
                         if(a0 \le 0 \mid | b0 \le 0)
                                                 \#Avoiding\ negative\ values
                                                 lnp < -\tilde{l}nf
                         else \{lnp \leftarrow lgamma(a0+b0) + lgamma(aN) + l
                         lgamma(b0+N-sum(y)) - lgamma(a0) - lgamma(b0) -
                         lgamma(a0+b0+N)
                         # log marginal likelihood
                         return(-lnp)
theta0 <- c(0.1, 0.1)
# Initial values
control <- list (maxit = 1000)
\# Number of iterations in optimization
EmpBay <- optim(theta0, LogMgLik, method = "BFGS",
control = control, hessian = TRUE, y = y)
# Optimization
EmpBay$convergence
# Checking convergence
EmpBay$value # Maximum
a0EB <- EmpBay$par[1]
# Prior shape using empirical Bayes
b0EB <- EmpBay$par[2]
# Prior scale using empirical Bayes
PriMeanEB <- a0EB /(a0EB + b0EB)
# Prior mean
PriVarEB <- (a0EB*b0EB)/(((a0EB+b0EB)^2)*(a0EB+b0EB+1)
# Prior variance
PpEB \leftarrow ProbBo(y, a0 = a0EB, b0 = b0EB)
  # This setting is using empirical Bayes.
PpEB
0.47
```

R code. Car claim, density plots

```
\# Density figures
# Density figures lambda <- seq(0.001, 1, 0.001) # Values of lambda VaguePrior <- dbeta(lambda, shape1 = a0, shape2 = b0) EBPrior <- dbeta(lambda, shape1 = a0EB, shape2 = b0EB) PosteriorV <- dbeta(lambda, shape1 = a0 + secup s
 # Likelihood function
Likelihood <- function(theta, y){
    LogL <- dbinom(y, 1, theta, log = TRUE)
    # LogL <- dbern(y, theta)
    Lik <- prod(exp(LogL))
    return(Lik)</pre>
                               return (Lik)
 }
 Liks <- sapply(lambda, function(par) {
 Likelihood(par, y = y)})
Sc <- max(PosteriorEB)/max(Liks)
 #Scale for displaying in figure
LiksScale <- Liks * Sc
  data <- data.frame(cbind(lambda, VaguePrior, EBPrior,
  PosteriorV, PosteriorEB, LiksScale))
 #Data frame
  require (ggplot2)
 # Cool figures
require(latex2exp)
  # LaTeX equations in figures
  require (ggpubr)
  # Multiple figures in one page
  fig1 <- ggplot(data = data, aes(lambda, VaguePrior)) +
 geom_line() + xlab(TeX("$p$")) + ylab("Density") + ggtitle("Prior:_Vague_Beta")
  fig2 <- ggplot(data = data, aes(lambda, EBPrior)) +
 geom_line() +
xlab(TeX("$p$")) + ylab("Density") +
ggtitle("Prior:_Empirical_Bayes_Beta")
  \label{eq:fig3} \textit{ $\leftarrow$ ggplot(data = data, aes(lambda, PosteriorV))$ + }
 geom_line() +
xlab(TeX("$p$")) + ylab("Density") +
  ggtitle ("Posterior: _Vague_Beta")
  fig4 <- ggplot(data = data, aes(lambda, PosteriorEB)) +
 geom_line() +
xlab(TeX("$p$")) + ylab("Density") +
ggtitle("Posterior:_Empirical_Bayes_Beta")
 \begin{array}{ll} FIG < & ggarrange\,(\,fig\,1\ ,\ fig\,2\ ,\ fig\,3\ ,\ fig\,4\ ,\\ n\,c\,o\,l \ = \ 2\ ,\ nrow \ = \ 2\,) \end{array}
  annotate_figure (FIG,
 top = text_grob("Vague_versus_Empirical_Bayes:
Beta-Bernoulli_model", color = "black",
face = "bold", size = 14))
```

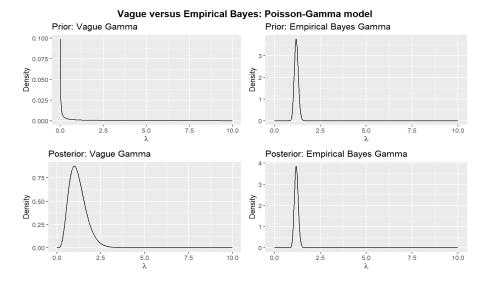


FIGURE 1.1 Vague versus Empirical Bayes: Bernoulli-Beta model.

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

```
# Prior, likelihood and posterior:
#Empirical Bayes Binonial-Beta model
dataNew <- data.frame(cbind(rep(lambda, 3),
c(EBPrior, PosteriorEB, LiksScale),
rep(1:3, each = 1000)))
#Data frame

colnames(dataNew) <- c("Lambda", "Density", "Factor")
dataNew$Factor <- factor(dataNew$Factor, levels=c("1", "3",
"2"), labels=c("Prior", "Likelihood", "Posterior"))

ggplot(data = dataNew, aes_string(x = "Lambda",
y = "Density", group = "Factor")) +
geom_line(aes(color = Factor)) +
xlab(TeX("$\\lambda$")) + ylab("Density") +
ggtitle("Prior,_likelihood_and_posterior:_Empirical_Bayes_Poisson-Gamma_model") +
guides(color=guide_legend(title="Information")) +
scale_color_manual(values = c("red", "yellow", "blue"))</pre>
```

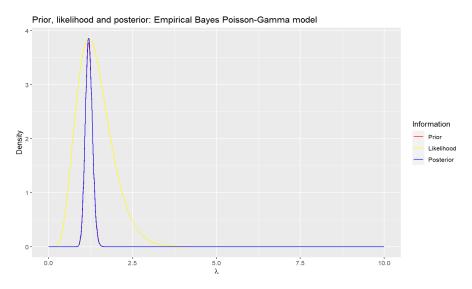


FIGURE 1.2
Prior, likelihood and posterior: Bernoulli-Beta model.

R code. Car claim, predictive probabilities plots

```
# Predictive distributions
require(TailRank)

PredDen <- function(y, y0, a0, b0){
        N <- length(y)
        aN <- a0 + sum(y) # Posterior shape parameter
        bN <- b0 + N - sum(y) # Posterior scale parameter
        Pr <- aN/(aN+bN)
        Probs <- dbinom(y0, 1, prob = Pr)
        return(Probs)
}
y0 <- 0:1
PredVague <- PredDen(y = y, y0 = y0, a0 = a0, b0 = b0)
PredEB <- PredDen(y = y, y0 = y0, a0 = a0EB, b0 = b0EB)
dataPred <- as.data.frame(cbind(y0, PredVague, PredEB))
colnames(dataPred) <- c("y0", "PredictiveVague",
"PredictiveEB")

ggplot(data = dataPred) +
geom_point(aes(y0, PredictiveVague, color = "red")) +
xlab(TeX("$y.0$")) + ylab("Density") +
ggtitle("Predictive_density:_Vague_and_Empirical
Bayes_priors")
+ geom_point(aes(y0, PredictiveEB, color = "yellow")) +
guides(color = guide_legend(title="Prior")) +
scale_color_manual(labels = c("Vague", "Empirical_Bayes"),
values = c("red", "yellow")) +
scale_x_continuous(breaks=seq(0,1,by=1))</pre>
```

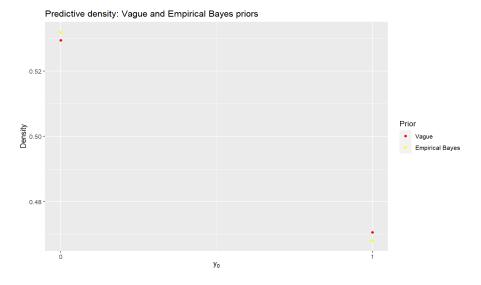


FIGURE 1.3
Predictive probabilities: Bernoulli-Beta model.

R code. Car claim, Bayesian model average

```
# Posterior odds: Vague vs Empirical Bayes
PO12 \leftarrow \exp(-\text{LogMgLik}(c(a0EB, b0EB), y = y))/
\exp(-\text{LogMgLik}(c(a0, b0), y = y))
PostProMEM \leftarrow PO12/(1 + PO12)
\# Posterior model probability Empirical Bayes PostProMEM
0.757
PostProbMV \leftarrow 1 - PostProMEM
# Posterior model probability vague prior
PostProbMV
0.242
# Bayesian model average (BMA)
PostMeanEB <- (a0EB + sum(y)) / (a0EB + b0EB + N)
# Posterior mean Empirical Bayes
PostMeanV <- (a0 + sum(y)) / (a0 + b0 + N)
# Posterior mean vague priors
BMAmean <- PostProMEM * PostMeanEB + PostProbMV * PostMeanV
# BMA nostErior mean
# BMA posterior mean
# Posterior variance vague prior
BMAVar < - PostProMEM * PostVarEB + PostProbMV * PostVarV
+ PostProMEM * (PostMeanEB - BMAmean)^2 + PostProbMV * (PostMeanV - BMAmean)^2
# BMA posterior variance
^{''}BMAPred <- PostProMEM * PredEB + PostProbMV * PredVague
dataPredBMA <- as.data.frame(cbind(y0, BMAPred))
colnames(dataPredBMA) <- c("y0", "PredictiveBMA")
ggplot(data = dataPredBMA) +
geom_point(aes(y0, PredictiveBMA, color = "red")) + xlab(TeX("$y_0$")) + ylab("Density") +
ggtitle ("Predictive_density: _BMA")
guides (color = guide_legend (title="BMA")) +
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

```
# Bayesian updating
BayUp <- function(y, lambda, a0, b0){
N <- length(y)
aN <- a0 + sum(y)
# Posterior shape parameter
bN <- b0 + N - sum(y)
# Posterior scale parameter
p <- dbeta (lambda, shape1 = aN, shape2 = bN)
# Posterior density
return(list(Post = p, a0New = aN, b0New = bN))
}

PostUp <- NULL
for(i in 1:N){
if (i = 1){
PostUpi <- BayUp(y[i], lambda, a0 = 1, b0 = 1)}
else{
PostUpi <- BayUp(y[i], lambda, a0 = PostUpi$b0New)
}
PostUp <- cbind(PostUp, PostUpi$Post)
}

DataUp <- data.frame(cbind(rep(lambda, 15), c(PostUp), rep(1:15, each = 1000)))
#Data frame
colnames(DataUp) <- c("Lambda", "Density", "Factor")
DataUp$Factor <- factor(DataUp$Factor, levels=c("1","2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "13", "14", "15")
labels=c("Iter_1", "Iter_2", "Iter_3", "Iter_4", "Iter_5", "Iter_6", "Iter_6", "Iter_7", "Iter_8", "Iter_9", "Iter_10", "Iter_11", "Iter_12", "Iter_11", "Iter_14", "Iter_15"))

ggplot(data = DataUp, aes_string(x = "Lambda", y = "Density", group = "Factor") + geom_line(aes(color = Factor)) + xlab(TeX("$p$")) + ylab("Density") + getile("Bayesian_updating:
_Beta_Binomial_model_with_vague_prior") + guides(color=guide_legend(title="Update"))
```

5. Show that given the loss function, $L(\theta, a) = |\theta - a|$, then $\delta(y)$ 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,

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

then,

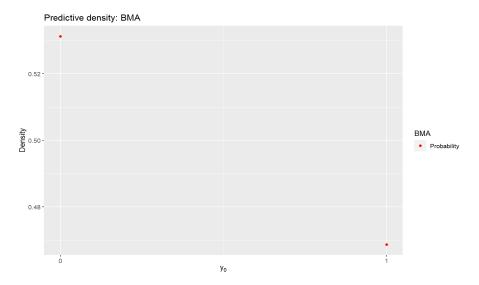


FIGURE 1.4 Predictive probabilities: Bernoulli-Beta Bayesian model average.

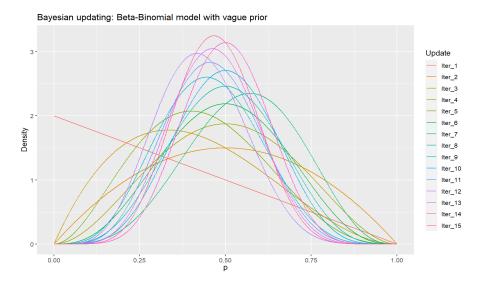


FIGURE 1.5
Predictive probabilities: Bernoulli-Beta Bayesian model updating.

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

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

Solutions of chapter 2 Conceptual differences: Bayesian and Frequentist approaches

2.1 Solutions Exercises

1. Jeffreys-Lindley's paradox

The **Jeffreys-Lindley's paradox** [1, 2] 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 $\mathcal{U}(0,1)$ under H_1 .
- Show that the *p*-value for this hypothesis test is equal to 0.023 using the normal approximation, $Y \sim \mathcal{N}(N \times \theta, N \times \theta \times (1 \theta))$.

Answer

• The marginal likelihood under the null hypothesis is $p(y|H_0) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = 1.95e - 04$ 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

24 Solutions of chapter 2 Conceptual differences: Bayesian and Frequentist approaches

$$p(y|H_1) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta$$

$$= \binom{n}{y} B(y+1, n-k+1)$$

$$= \frac{\Gamma(N+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(y+1)\Gamma(N-y+1)}{\Gamma(N+2)}$$

$$= \frac{N!}{(N+1)!}$$

$$= \frac{1}{N+1}$$

$$= 1.016e - 05$$

Then, $PO_{01} = \frac{1.95e - 04}{1.016e - 05} = 19.19$, 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,

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

= 0.0235,

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

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

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

$$\begin{split} p(\mathbf{y}|\mathcal{M}_1) &= \int_{-\infty}^{\infty} \int_{0}^{\infty} (2\pi)^{-N/2} \sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2\right\} \frac{1}{\sigma} d\sigma d\mu \\ &= (2\pi)^{-N/2} \int_{-\infty}^{\infty} \int_{0}^{\infty} \sigma^{-(N+1)} \exp\left\{-\frac{N}{2\sigma^2} \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}\right\} d\sigma d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \int_{-\infty}^{\infty} \left[\sum_{i=1}^{N} (y_i - \mu)^2\right]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \int_{-\infty}^{\infty} \left[\sum_{i=1}^{N} [(y_i - \bar{y}) - (\mu - \bar{y})]^2\right]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \int_{-\infty}^{\infty} \left[\alpha_n \hat{\sigma}^2 + N(\mu - \bar{y})^2\right]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \left(\frac{\alpha_n \hat{\sigma}^2}{\alpha_n \hat{\sigma}^2}\right)^{-N/2} \int_{-\infty}^{\infty} \left[\alpha_n \hat{\sigma}^2 + N(\mu - \bar{y})^2\right]^{-N/2} d\mu \\ &= (2\pi)^{-N/2} \frac{\Gamma(N/2)}{2} 2^{N/2} \left(\alpha_n \hat{\sigma}^2\right)^{-N/2} \int_{-\infty}^{\infty} \left[1 + \frac{N(\mu - \bar{y})^2}{\alpha_n \hat{\sigma}^2}\right]^{-N/2} d\mu \\ &= \frac{\pi^{-N/2}}{2} \Gamma(N/2) 2^{N/2} \left(\frac{1}{\alpha_n \hat{\sigma}^2}\right)^{N/2} \left(\frac{N}{\alpha_n \hat{\sigma}^2}\right)^{-1/2} \frac{\Gamma(1/2)\Gamma(\alpha_n/2)}{\Gamma((\alpha_N + 1)/2)}. \end{split}$$

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 [5].

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

The third line takes into account that the integral in the second line is the kernel of an inverted-gamma distribution [5].

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

In addition,

$$PO_{01} = \frac{\Gamma((\alpha_{n}+1)/2)}{\Gamma(1/2)\Gamma(\alpha_{N}/2)} (\alpha_{n}\hat{\sigma}^{2}/N)^{-1/2} \left[1 + \frac{(\mu_{0}-\bar{y})^{2}}{\alpha_{n}\hat{\sigma}^{2}/N} \right]^{-\left(\frac{\alpha_{n}+1}{2}\right)}$$

$$= \frac{\Gamma((\alpha_{n}+1)/2)}{\Gamma(1/2)\Gamma(\alpha_{N}/2)} (\alpha_{n}\hat{\sigma}^{2}/N)^{-1/2} \left[1 + \frac{1}{\alpha_{n}} \left(\frac{\mu_{0}-\bar{y}}{\hat{\sigma}/\sqrt{N}} \right)^{2} \right]^{-\left(\frac{\alpha_{n}+1}{2}\right)}$$

$$= \frac{\Gamma((\alpha_{n}+1)/2)}{\Gamma(1/2)\Gamma(\alpha_{N}/2)} (\alpha_{n}\hat{\sigma}^{2}/N)^{-1/2} \left[1 + \frac{1}{\alpha_{n}} t^{2} \right]^{-\left(\frac{\alpha_{n}+1}{2}\right)},$$

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. 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

Solutions of chapter 2 Cornerstone models: Conjugate families

3.1 Solutions 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

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{N} \binom{n_i}{y_i} \theta^{y_i} (1-\theta)^{n_i-y_i}$$

$$= \prod_{i=1}^{N} \binom{n_i}{y_i} \theta^{\sum_{i=1}^{N} y_i} (1-\theta)^{\sum_{i=1}^{N} n_i - \sum_{i=1}^{N} y_i}$$

$$= \prod_{i=1}^{N} \binom{n_i}{y_i} \exp\left\{ \sum_{i=1}^{N} y_i \log\left(\frac{\theta}{1-\theta}\right) + \sum_{i=1}^{N} n_i \log(1-\theta) \right\}$$

$$= \prod_{i=1}^{N} \binom{n_i}{y_i} (1-\theta)^{\sum_{i=1}^{N} n_i} \exp\left\{ \sum_{i=1}^{N} y_i \log\left(\frac{\theta}{1-\theta}\right) \right\},$$

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

$$\pi(\theta) \propto (1 - \theta)^{\mathbf{B}_0} \exp\left\{a_0 \log\left(\frac{\theta}{1 - \theta}\right)\right\}$$
$$= \theta^{a_0} (1 - \theta)^{\mathbf{B}_0 - a_0}$$
$$= \theta^{\alpha_0 - 1} (1 - \theta)^{\beta_0 - 1},$$

where $\alpha_0 = a_0 + 1$ and $\beta_0 = \mathbf{B}_0 - a_0 + 1$. This is the kernel of a beta distribution. Thus, the posterior distribution is

$$\pi(\theta|\mathbf{y}) \propto \theta^{\alpha_0 - 1} (1 - \theta)^{\beta_0 - 1} \times \theta^{\sum_{i=1}^N y_i} (1 - \theta)^{\sum_{i=1}^N n_i - \sum_{i=1}^N y_i}$$

$$= \theta^{\alpha_0 + \sum_{i=1}^N y_i - 1} (1 - \theta)^{\beta_0 + \sum_{i=1}^N n_i - \sum_{i=1}^N y_i - 1}$$

$$= \theta^{\alpha_n - 1} (1 - \theta)^{\beta_n - 1},$$

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

$$p(\mathbf{y}) = \int_{0}^{1} \frac{\theta^{\alpha_{0}-1} (1-\theta)^{\beta_{0}-1}}{B(\alpha_{0},\beta_{0})} \times \prod_{i=1}^{N} \binom{n_{i}}{y_{i}} \theta^{\sum_{i=1}^{N} y_{i}} (1-\theta)^{\sum_{i=1}^{N} n_{i} - \sum_{i=1}^{N} y_{i}} d\theta$$

$$= \frac{\prod_{i=1}^{N} \binom{n_{i}}{y_{i}}}{B(\alpha_{0},\beta_{0})} \int_{0}^{1} \theta^{\alpha_{0} + \sum_{i=1}^{N} y_{i} - 1} (1-\theta) \beta_{0} \sum_{i=1}^{N} n_{i} - \sum_{i=1}^{N} y_{i} - 1 d\theta$$

$$= \frac{\prod_{i=1}^{N} \binom{n_{i}}{y_{i}} B(\alpha_{n},\beta_{n})}{B(\alpha_{0},\beta_{0})}.$$

The third line due to having the kernel of a Beta distribution. Finally, the predictive distribution is

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

$$= \frac{\binom{n_{y_0}}{y_0}}{B(\alpha_n, \beta_n)} \int_0^1 \theta^{\alpha_n+y_0-1} (1-\theta)^{\beta_n+n_{y_0}-y_0-1} d\theta$$

$$= \binom{n_{y_0}}{y_0} \frac{B(\alpha_n + y_0, \beta_n + n_{y_0} - y_0)}{B(\alpha_n, \beta_n)},$$

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

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

Finally, the predictive distribution is

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

This is a Lomax distribution.

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

Answer

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

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

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

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

Answer

$$p(\mathbf{Y}) = \int_{\mathcal{R}^{p}} \int_{\mathcal{S}} (2\pi)^{-pN/2} |\mathbf{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2} tr[(\mathbf{S} + N(\mu - \hat{\mu})(\mu - \hat{\mu})^{\top})\mathbf{\Sigma}^{-1}]\right\}$$

$$\times (2\pi)^{-p/2} \beta_{0}^{p/2} |\mathbf{\Sigma}|^{-1/2} \exp\left\{-\frac{\beta_{0}}{2} tr[(\mu - \mu_{0})(\mu - \mu_{0})^{\top}\mathbf{\Sigma}^{-1}]\right\}$$

$$\times |\mathbf{\Sigma}|^{-(\alpha_{0}+p+1)/2} \frac{2^{-\alpha_{0}p/2} |\mathbf{\Psi}_{0}|^{\alpha_{0}/2}}{\Gamma_{p}(\alpha_{0}/2)} \exp\left\{-\frac{1}{2} tr(\mathbf{\Psi}_{0}\mathbf{\Sigma}^{-1})\right\} d\mathbf{\Sigma} d\mu$$

$$= \frac{(2\pi)^{-frac12(pN+p)} |\mathbf{\Psi}_{0}|^{\alpha_{0}/2} \beta_{0}^{p/2} 2^{-\alpha_{0}p/2}}{\Gamma_{p}(\alpha_{0}/2)} \int_{\mathcal{R}^{p}} \int_{\mathcal{S}} |\mathbf{\Sigma}|^{-\frac{1}{2}(N+1+\alpha_{0}+p+1)}$$

$$\times \exp\left\{-\frac{1}{2} tr[(\mathbf{S} + N(\mu - \hat{\mu})(\mu - \hat{\mu})^{\top} + \beta_{0}(\mu - \mu_{0})(\mu - \mu_{0})^{\top} + \mathbf{\Psi}_{0})\mathbf{\Sigma}^{-1}]\right\} d\mathbf{\Sigma} d\mu.$$

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

$$p(\mathbf{Y}) = \frac{\Gamma_{p} \left(\frac{N+1+\alpha_{0}}{2}\right) |\mathbf{\Psi}_{0}|^{\alpha_{0}/2} \beta_{0}^{p/2}}{\Gamma_{p} (\alpha_{0}/2) \pi^{p(N+1)/2}}$$

$$\times \int_{\mathcal{R}^{p}} |\mathbf{S} + \mathbf{\Psi}_{0} + (N+\beta_{0}) (\mu - \mu_{n}) (\mu - \mu_{n})^{\top}$$

$$+ N\beta_{0}/(N+\beta_{0}) (\hat{\mu} - \mu_{0}) (\hat{\mu} - \mu_{0})^{\top} |d\mu$$

$$= \frac{\Gamma_{p} \left(\frac{N+1+\alpha_{0}}{2}\right) |\mathbf{\Psi}_{0}|^{\alpha_{0}/2} \beta_{0}^{p/2}}{\Gamma_{p} (\alpha_{0}/2) \pi^{p(N+1)/2}}$$

$$\times \int_{\mathcal{R}^{p}} |\mathbf{\Psi}_{n}| |1 + \beta_{n} (\mu - \mu_{n}) \mathbf{\Psi}_{n}^{-1} (\mu - \mu_{n})^{\top} |^{-\frac{1}{2} (\alpha_{n}+1)} d\mu$$

$$= \frac{\Gamma_{p} \left(\frac{\alpha_{n}+1}{2}\right) |\mathbf{\Psi}_{0}|^{\alpha_{0}/2} \beta_{0}^{p/2}}{\Gamma_{p} (\alpha_{0}/2) \pi^{p(N+1)/2}} |\mathbf{\Psi}_{n}|^{-\frac{1}{2} (\alpha_{n}+1)}$$

$$\times \int_{\mathcal{R}^{p}} [1 + \beta_{n} (\mu - \mu_{n})^{\top} \mathbf{\Psi}_{n}^{-1} (\mu - \mu_{n})]^{-\frac{1}{2} (\alpha_{n}+1)} d\mu.$$

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 Student's t distribution [3]. Then,

$$p(\mathbf{Y}) = \frac{\Gamma_{p} \left(\frac{\alpha_{n}+1}{2}\right) |\mathbf{\Psi}_{0}|^{\alpha_{0}/2} \beta_{0}^{p/2}}{\Gamma_{p} (\alpha_{0}/2) \pi^{p(N+1)/2}} |\mathbf{\Psi}_{n}|^{-\frac{1}{2}(\alpha_{n}+1)}$$

$$\times \int_{\mathcal{R}^{p}} \left[1 + \frac{1}{\alpha_{n}+1-p} (\mu - \mu_{n})^{\top} \left(\frac{\mathbf{\Psi}_{n}}{\beta_{n} (\alpha_{n}+1-p)} \right)^{-1} (\mu - \mu_{n}) \right]^{-\frac{1}{2}(\alpha_{n}+1-p+p)} d\mu$$

$$= \frac{\Gamma_{p} \left(\frac{\alpha_{n}+1}{2}\right) \Gamma_{p} \left(\frac{\alpha_{n}+1-p}{2}\right) |\mathbf{\Psi}_{0}|^{\alpha_{0}/2} \beta_{0}^{p/2} (\alpha_{n}+1-p)^{p/2} \pi^{p/2} |\mathbf{\Psi}_{n}|^{-\frac{1}{2}(\alpha_{n}+1)}}{\Gamma_{p} (\alpha_{0}/2) \pi^{p(N+1)/2} \Gamma_{p} \left(\frac{\alpha_{n}+1-p+p}{2}\right) \left(\frac{\mathbf{\Psi}_{n}}{\alpha_{n}+1-p}\right)^{-1/2}}$$

$$= \frac{\Gamma_{p} \left(\frac{v_{n}}{2}\right)}{\Gamma_{p} \left(\frac{\alpha_{0}}{2}\right)} \frac{|\mathbf{\Psi}_{0}|^{\alpha_{0}/2}}{|\mathbf{\Psi}_{n}|^{\alpha_{n}/2}} \left(\frac{\beta_{0}}{\beta_{n}}\right)^{p/2} (2\pi)^{-Np/2},$$

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

6.

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

Answer

$$p(\mathbf{Y}_0|\mathbf{Y}) \propto \int_{\mathcal{R}^p} \int_{\mathcal{S}} |\mathbf{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2} tr[(\mathbf{y}_0 - \mu)(\mathbf{y}_0 - \mu)^\top \mathbf{\Sigma}^{-1}]\right\}$$

$$\times |\mathbf{\Sigma}|^{-1/2} \exp\left\{-\frac{\beta_n}{2} tr[(\mu - \mu_n)(\mu - \mu_n)^\top \mathbf{\Sigma}^{-1}]\right\}$$

$$\times |\mathbf{\Sigma}|^{-(\alpha_n + p + 1)/2} \exp\left\{-\frac{1}{2} tr(\mathbf{\Psi}_n \mathbf{\Sigma}^{-1})\right\} d\mathbf{\Sigma} d\mu$$

$$\propto \int_{\mathcal{R}^p} |(\mathbf{y}_0 - \mu)(\mathbf{y}_0 - \mu)^\top + (\mu - \mu_n)(\mu - \mu_n)^\top + \mathbf{\Psi}_n|^{-(\alpha_n + 2)/2} d\mu.$$

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

Taking into account that

$$(\mathbf{y}_{0} - \mu)(\mathbf{y}_{0} - \mu)^{\top} + (\mu - \mu_{n})(\mu - \mu_{n})^{\top} = (1 + \beta_{n}) \left(\mu - \frac{(\mathbf{y}_{0} + \beta_{n}\mu_{n})}{1 + \beta_{n}}\right) \left(\mu - \frac{(\mathbf{y}_{0} + \beta_{n}\mu_{n})}{1 + \beta_{n}}\right)^{\top} + \frac{\beta_{n}}{1 + \beta_{n}}(\mathbf{y}_{0} - \mu_{n})(\mathbf{y}_{0} - \mu_{n})^{\top}.$$

Then,

$$\begin{aligned} p(\mathbf{Y}_{0}|\mathbf{Y}) &\propto \int_{\mathcal{R}^{p}} \left| (\mathbf{y}_{0} - \mu)(\mathbf{y}_{0} - \mu)^{\top} + (\mu - \mu_{n})(\mu - \mu_{n})^{\top} + \mathbf{\Psi}_{n} \right|^{-(\alpha_{n} + 2)/2} d\mu \\ &= \int_{\mathcal{R}^{p}} \left| (1 + \beta_{n}) \left(\mu - \frac{(\mathbf{y}_{0} + \beta_{n}\mu_{n})}{1 + \beta_{n}} \right) \left(\mu - \frac{(\mathbf{y}_{0} + \beta_{n}\mu_{n})}{1 + \beta_{n}} \right)^{\top} \right. \\ &+ \frac{\beta_{n}}{1 + \beta_{n}} (\mathbf{y}_{0} - \mu_{n})(\mathbf{y}_{0} - \mu_{n})^{\top} + \mathbf{\Psi}_{n} \right|^{-(\alpha_{n} + 2)/2} d\mu \\ &= \int_{\mathcal{R}^{p}} \left[\left| \mathbf{\Psi}_{n} + \frac{\beta_{n}}{1 + \beta_{n}} (\mathbf{y}_{0} - \mu_{n})(\mathbf{y}_{0} - \mu_{n})^{\top} \right| \right. \\ &\left. \left. \left| 1 + (1 + \beta_{n}) \left(\mu - \frac{(\mathbf{y}_{0} + \beta_{n}\mu_{n})}{1 + \beta_{n}} \right)^{\top} \frac{1}{\alpha_{n} + 2 - p} \left(\frac{\mathbf{\Lambda}_{n}}{\alpha_{n} + 2 - p} \right)^{-1} \left(\mu - \frac{(\mathbf{y}_{0} + \beta_{n}\mu_{n})}{1 + \beta_{n}} \right) \right| \right]^{-(\alpha_{n} + 2 - p + p)/2} d\mu \\ &\propto \left| \mathbf{\Psi}_{n} + \frac{\beta_{n}}{1 + \beta_{n}} (\mathbf{y}_{0} - \mu_{n})(\mathbf{y}_{0} - \mu_{n})^{\top} \right|^{-(\alpha_{n} + 2)/2} \\ &\times \left| \mathbf{\Psi}_{n} + \frac{\beta_{n}}{1 + \beta_{n}} (\mathbf{y}_{0} - \mu_{n})(\mathbf{y}_{0} - \mu_{n})^{\top} \right|^{1/2} \\ &= \left| \mathbf{\Psi}_{n} + \frac{\beta_{n}}{1 + \beta_{n}} (\mathbf{y}_{0} - \mu_{n})(\mathbf{y}_{0} - \mu_{n})^{\top} \right|^{-(\alpha_{n} + 1)/2} \\ &\propto \left[1 + (\mathbf{y}_{0} - \mu_{n})^{\top} \frac{1}{\alpha_{n} + 1 - p} \left(\frac{\mathbf{\Psi}_{n}(1 + \beta_{n})}{(\alpha_{n} + 1 - p)\beta_{n}} \right)^{-1} (\mathbf{y}_{0} - \mu_{n}) \right]^{-(\alpha_{n} + 1 - p + p)} . \end{aligned}$$

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 Student's t distribution.

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

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

$$\pi(\mathbf{Y}_{0}|\mathbf{Y}) \propto \int_{\mathcal{S}} \int_{\mathcal{B}} \left\{ |\mathbf{\Sigma}|^{-N_{0}/2} \exp\left\{-\frac{1}{2}tr[(\mathbf{Y}_{0} - \mathbf{X}_{0}\mathbf{B})^{\top}(\mathbf{Y}_{0} - \mathbf{X}_{0}\mathbf{B})\mathbf{\Sigma}^{-1}]\right\} \right.$$

$$\times |\mathbf{\Sigma}|^{-K/2} \exp\left\{-\frac{1}{2}tr[(\mathbf{B} - \mathbf{B}_{n})^{\top}\mathbf{V}_{n}^{-1}(\mathbf{B} - \mathbf{B}_{n})\mathbf{\Sigma}^{-1}]\right\}$$

$$\times |\mathbf{\Sigma}|^{-(\alpha_{n}+M+1)/2} \exp\left\{-\frac{1}{2}tr[\mathbf{\Psi}_{n}\mathbf{\Sigma}^{-1}]\right\} d\mathbf{B}d\mathbf{\Sigma}$$

$$= \int_{\mathcal{S}} \int_{\mathcal{B}} \left\{ |\mathbf{\Sigma}|^{-(N_{0}+K+\alpha_{n}+M+1)/2} \exp\left\{-\frac{1}{2}tr[((\mathbf{Y}_{0} - \mathbf{X}_{0}\mathbf{B})^{\top}(\mathbf{Y}_{0} - \mathbf{X}_{0}\mathbf{B}) + (\mathbf{B} - \mathbf{B}_{n})^{\top}\mathbf{V}_{n}^{-1}(\mathbf{B} - \mathbf{B}_{n}) + \mathbf{\Psi}_{n})\mathbf{\Sigma}^{-1}]\right\} d\mathbf{B}d\mathbf{\Sigma}.$$

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,

$$\pi(\mathbf{Y}_0|\mathbf{Y}) \propto \int_{\mathcal{S}} |\mathbf{\Sigma}|^{-(N_0 + K + \alpha_n + M + 1)/2}$$

$$\times \exp\left\{-\frac{1}{2}tr[(\mathbf{\Psi}_n + \mathbf{B}_n^{\top}\mathbf{V}_n^{-1}\mathbf{B}_n + \mathbf{Y}_0^{\top}\mathbf{Y}_0 - \mathbf{B}_*^{\top}\mathbf{M}\mathbf{B}_*)\mathbf{\Sigma}^{-1}]\right\}$$

$$\times \int_{\mathcal{B}} \exp\left\{-\frac{1}{2}tr[(\mathbf{B} - \mathbf{B}_*)^{\top}\mathbf{M}(\mathbf{B} - \mathbf{B}_*)\mathbf{\Sigma}^{-1}]\right\} d\mathbf{B}d\mathbf{\Sigma}.$$

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

$$\pi(\mathbf{Y}_0|\mathbf{Y}) \propto \int_{\mathcal{S}} |\mathbf{\Sigma}|^{-(N_0 + \alpha_n + M + 1)/2}$$

$$\times \exp\left\{-\frac{1}{2}tr[(\mathbf{\Psi}_n + \mathbf{B}_n^{\top}\mathbf{V}_n^{-1}\mathbf{B}_n + \mathbf{Y}_0^{\top}\mathbf{Y}_0 - \mathbf{B}_*^{\top}\mathbf{M}\mathbf{B}_*)\mathbf{\Sigma}^{-1}]\right\} d\mathbf{\Sigma}$$

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

$$\pi(\mathbf{Y}_0|\mathbf{Y}) \propto \left|\mathbf{\Psi}_n + \mathbf{B}_n^{ op} \mathbf{V}_n^{-1} \mathbf{B}_n + \mathbf{Y}_0^{ op} \mathbf{Y}_0 - \mathbf{B}_*^{ op} \mathbf{M} \mathbf{B}_* \right|^{-(N_0 + lpha_n)/2}$$

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,

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

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 $\mathbf{\Psi}_n$ and $\mathbf{C}^{-1} = \mathbf{I}_{N_0} + \mathbf{X}_0 \mathbf{V}_n \mathbf{X}_0^{\top}$.

8. Show that $\delta_n = \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\beta})^{\top}(\mathbf{y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - \beta_0)^{\top}((\mathbf{X}^{\top}\mathbf{X})^{-1} + \mathbf{B}_0)^{-1}(\hat{\beta} - \beta_0)$ in the linear regression model, and that $\mathbf{\Psi}_n = \mathbf{\Psi}_0 + \mathbf{S} + (\hat{\mathbf{B}} - \mathbf{B}_0)^{\top}\mathbf{V}_n(\hat{\mathbf{B}} - \mathbf{B}_0)$ in the linear multivariate regression model.

Answer

Taking into account that

$$\begin{split} \delta^* &= \delta_0 + \mathbf{y}^\top \mathbf{y} + \beta_0^\top \mathbf{B}_0^{-1} \beta_0 - \beta_n^\top \mathbf{B}_n^{-1} \beta_n \\ &= \delta_0 + \mathbf{y}^\top \mathbf{y} + \beta_0^\top \mathbf{B}_0^{-1} \beta_0 - (\mathbf{B}_0^{-1} \beta_0 + \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}})^\top \mathbf{B}_n (\mathbf{B}_0^{-1} \beta_0 + \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}) \\ &= \delta_0 + \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} - 2 \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{B}_0^{-1} \beta_0 + \beta_0^\top (\mathbf{B}_0^{-1} - \mathbf{B}_0^{-1} \mathbf{B}_n \mathbf{B}_0^{-1}) \beta_0 \\ &- \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \delta_0 + \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top (\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} \\ &- 2 \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{B}_n \mathbf{B}_0^{-1} \beta_0 + \beta_0^\top (\mathbf{B}_0^{-1} - \mathbf{B}_0^{-1} \mathbf{B}_n \mathbf{B}_0^{-1}) \beta_0. \end{split}$$

Observe that

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

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

The following matrix identities are useful [4]:

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

and

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

Using these identities,

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

Then,

$$\begin{split} \delta^* &= \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} \hat{\boldsymbol{\beta}} \\ &- 2\hat{\boldsymbol{\beta}} [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} \beta_0 + \beta_0^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} \beta_0 \\ &= \delta_0 + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &+ (\hat{\boldsymbol{\beta}} - \beta_0)^\top [(\mathbf{X}^\top \mathbf{X})^{-1} + \mathbf{B}_0]^{-1} (\hat{\boldsymbol{\beta}} - \beta_0). \end{split}$$

In a similar way for the second part,

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

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 $\hat{\mathbf{B}} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{Y}$.

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

Answer

Taking into account that $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{A}^{-1}$ [4], 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

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

Let's show that $\beta_{**} = \mathbf{X}_0 \beta_n$,

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

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.

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

Answer

$$(\mathbf{Y} - \mathbf{X}\mathbf{B})^{\top}(\mathbf{Y} - \mathbf{X}\mathbf{B}) = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} + \mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B})^{\top}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} + \mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B})$$

$$= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^{\top}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) + 2(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^{\top}(\mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B})$$

$$+ (\mathbf{X}\mathbf{B} - \mathbf{X}\hat{\mathbf{B}})^{\top}(\mathbf{X}\mathbf{B} - \mathbf{X}\hat{\mathbf{B}})$$

$$= \mathbf{S} + (\mathbf{B} - \hat{\mathbf{B}})^{\top}\mathbf{X}^{\top}\mathbf{X}(\mathbf{B} - \hat{\mathbf{B}}),$$

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 $\mathbf{X}^{\top}\hat{\mathbf{U}} = \mathbf{0}$.

Bibliography

- [1] H. Jeffreys. Theory of Probability. Oxford University Press, London, 1961.
- [2] Dennis V Lindley. A statistical paradox. $Biometrika,\ 44(1/2):187-192,\ 1957.$
- [3] Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. def, $1(2\sigma 2):16$, 2007.
- [4] A. F. M. Smith. A General Bayesian Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological).*, 35(1):67–75, 1973.
- [5] A. Zellner. Introduction to Bayesian inference in econometrics. John Wiley & Sons Inc., 1996.