Half Title

Title Page

LOC Page

To my parents, Nancy and Orlando.

Contents

Foreword

Preface

Contributors

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

We introduce formal concepts in Bayesian inference starting with the Bayes' rule, all its components with their formal definitions and basic examples. In addition, we present some nice features of Bayesian inference such as Bayesian updating, and asymptotic sampling properties, and the basics of Bayesian inference based on decision theory under uncertainty, presenting important concepts like loss function, risk function and optimal rules.

1.1 The Bayes' rule

As expected the point of departure to perform Bayesian inference is the Bayes' rule,[1] which is the Bayes' solution to the inverse probability of causes, this rule combines prior beliefs with objective probabilities based on repeatable experiments. In this way, we can move from observations to probable causes.

Formally, the conditional probability of A_i given B is equal to the conditional probability of B given A_i times the marginal probability of A_i over the marginal probability of B,

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i, B)}{P(B)} \\ &= \frac{P(B|A_i) \times P(A_i)}{P(B)}, \end{aligned} \tag{1.1}$$

where by the law of total probability $P(B) = \sum_i P(B|A_i)P(A_i) \neq 0$, $\{A_i, i = 1, 2, \dots\}$ is a finite or countably infinite partition of a sample space.

In the Bayesian framework, B is sample information that updates a probabilistic statement about an unknown object A_i following probability rules. This is done by means of the Bayes' rule using prior "beliefs" about A_i, that is, $P(A_i)$, sample information relating B to the particular state of the nature A_i through a probabilistic statement, $P(B|A_i)$, and the probability of observing that specific sample information $P(B)$.

[1] Observe that I use the term "Bayes' rule" rather than "Bayes' theorem". It was Laplace [41] who actually generalized the Bayes' theorem [4]. His generalization is named the Bayes' rule.

Let's see a simple example, *the base rate fallacy*:

Assume that the sample information comes from a positive result from a test whose true positive rate (sensitivity) is 98%, $P(+|\text{disease}) = 0.98$. On the other hand, the prior information regarding being infected with this disease comes from a base incidence rate that is equal to 0.002, that is $P(\text{disease}) = 0.002$. Then, *what is the probability of being actually infected?*

This is an example of *the base rate fallacy*, where having a positive test result from a disease whose base incidence rate is tiny gives a low probability of actually having the disease.

The key to answer the question is based on understanding the difference between the probability of having the disease given a positive result, $P(\text{disease}|+)$, versus the probability of a positive result given the disease, $P(+|\text{disease})$. The former is the important result, and the Bayes' rule help us to get the answer. Using the Bayes' rule (equation 1.1):

$$\begin{aligned} P(\text{disease}|+) &= \frac{P(+|\text{disease}) \times P(\text{disease})}{P(+)} \\ &= \frac{0.98 \times 0.002}{0.98 \times 0.002 + (1 - 0.98) \times (1 - 0.002)} \\ &= 0.09, \end{aligned}$$

where $P(+) = P(+|\text{disease}) \times P(\text{disease}) + P(+|\neg\text{disease}) \times P(\neg\text{disease})$.[2]

R code. the base rate fallacy

```
PD <- 0.002 # Probability of disease
PPD <- 0.98 # True positive (Sensitivity)
PDP <- PD * PPD / (PD * PPD + (1 - PD)*(1 - PPD))
```

We observe that despite of having a positive result, the probability of having the disease is low. This due to the base rate being tiny.

Another interesting example, which is at the heart of the origin of the Bayes' theorem [4], is related to the existence of God [60]. The Section X of David Hume's "An Inquiry concerning Human Understanding, 1748" is named *Of Miracles*. There, Hume argues that when someone claims to have seen a miracle, this is poor evidence it actually happened, since it goes against what we see every day. Then, Richard Price, who actually finished and published "An essay towards solving a problem in the doctrine of chances" in 1763 after

[2] $\neg$ is the negation symbol. In addition, we have that $P(B|A) = 1 - P(B|A^c)$ in this example, where A^c is the complement of A. However, this is not true in general.

Bayes died in 1761, argues against Hume saying that there is a huge difference between *impossibility* as used commonly in conversation and *physical impossibility*. Price used an example of a dice with a million sides, where the former is getting a particular side when throwing this dice, and the latter is getting a side that does not exist. In millions throws, the latter case never would occur, but the former eventually would.

Let's say that there are two cases of resurrection (Res), Jesus Christ and Elvis, and the total number of people who have ever lived is 108.5 billion,[3] then the prior base rate is $2/(108.5 \times 10^9)$. On the other hand, the sample information comes from a very reliable witness whose true positive rate is 0.9999999. Then, *what is the probability of this miracle?*[4]

Using the Bayes' rule:

$$\begin{aligned} P(\text{Res}|\text{Witness}) &= \frac{P(\text{Witness}|\text{Res}) \times P(\text{Res})}{P(\text{Witness})} \\ &= \frac{2/(108.5 * 10^9) \times 0.99999}{2/(108.5 * 10^9) \times 0.99999 + (1 - 2/(108.5 * 10^9)) \times (1 - 0.99999)} \\ &= 0.000184297806959661 \end{aligned}$$

where $P(\text{Witness}) = P(\text{Witness}|\text{Res}) \times P(\text{Res}) + (1 - P(\text{Witness}|\text{Res})) \times (1 - P(\text{Res}))$.

R code. Of Miracles

```
# Probability of resurrection
PR <- 2/(108.5 * 10^9)
PWR <- 0.9999999 # True positive rate
PRW <- PR * PWR / (PR * PWR + (1 - PR)*(1 - PWR))
```

Observe that we can get a conditional version of the Bayes' rule. Let's have two conditioning events B and C, then equation 1.1 becomes

$$\begin{aligned} P(A_i|B,C) &= \frac{P(A_i,B,C)}{P(B,C)} \\ &= \frac{P(B|A_i,C) \times P(A_i|C) \times P(C)}{P(B|C)P(C)}. \end{aligned} \tag{1.2}$$

Let's use one of the most intriguing statistical puzzles, *the Monty Hall problem*, to illustrate how to use equation 1.2 [57, 58]. This was the situation

[3] https://www.wolframalpha.com/input/?i=number+of+people+who+have+ever+lived+on+Earth

[4] https://www.r-bloggers.com/2019/04/base-rate-fallacy-or-why-no-one-is-justified-to-believe-that-jesus-rose/

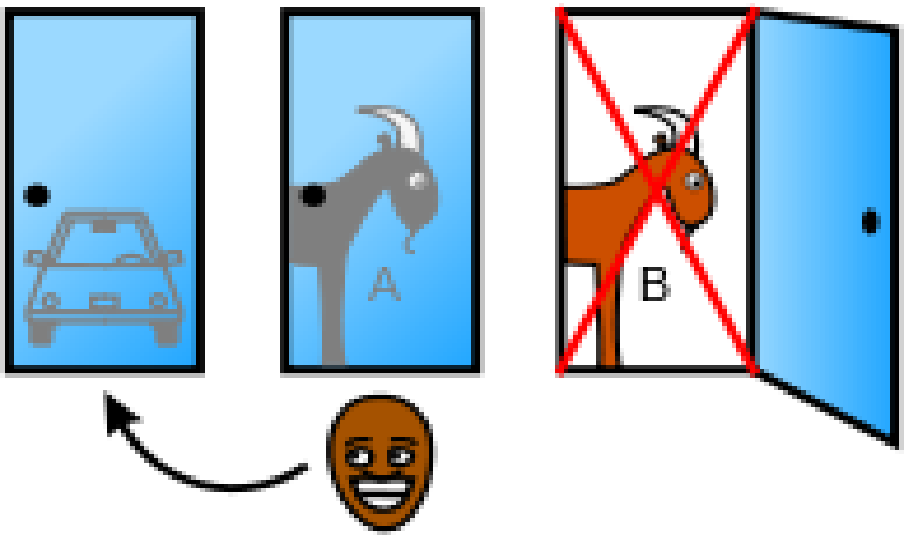

FIGURE 1.1
The Monty Hall problem.

faced by a contestant in the American television game show *Let's Make a Deal*. There, the contestant was asked to choose a door where behind one door there is a car, and behind the others, goats. Let's say that the contestant picks door No. 1, and the host (Monty Hall), who knows what is behind each door, opens door No. 3, where there is a goat (see Figure 1.1). Then, the host asks the tricky question to the contestant, *do you want to pick door No. 2?*

Let's name P_i the event **contestant picks door No.** i, which stays close, H_i the event **host picks door No.** i, which is open, and there is a goat, and C_i the event **car is behind door No.** i. In this particular setting, the contestant is interested in the probability of the event $P(C_2|H_3, P_1)$. A naive answer would be that it is irrelevant as initially $P(C_i) = 1/3,\ i = 1, 2, 3$, and now $P(C_i|H_3) = 1/2,\ i = 1, 2$ as the host opened door No. 3. So, why bothering changing the initial guess if the odds are the same (1:1)? The important point here is that the host knows what is behind each door and randomly picks a door given contestant choice. That is, $P(H_3|C_3, P_1) = 0$, $P(H_3|C_2, P_1) = 1$ and $P(H_3|C_1, P_1) = 1/2$. Then, using equation 1.2

$$
\begin{aligned}
P(C_2|H_3, P_1) &= \frac{P(C_2, H_3, P_1)}{P(H_3, P_1)} \\
&= \frac{P(H_3|C_2, P_1)P(C_2|P_1)P(P_1)}{P(H_3|P_1) \times P(P_1)} \\
&= \frac{P(H_3|C_2, P_1)P(C_2)}{P(H_3|P_1)} \\
&= \frac{1 \times 1/3}{1/2},
\end{aligned}
$$

where the third equation uses the fact that C_i and P_i are independent events, and $P(H_3|P_1) = 1/2$ due to this depending just on P_1 (not on C_2).

Therefore, changing the initial decision increases the probability of getting the car from 1/3 to 2/3!

Let's see a simulation exercise to check this answer:

R code. The Monty Hall problem

```
set.seed(0101) # Set simulation seed
S <- 100000 # Simulations
Game <- function(switch = 0){
        # switch = 0 is not change
        # switch = 1 is to change
        opts <- 1:3
        car <- sample(opts, 1) # car location
        guess1 <- sample(opts, 1) # Initial guess

        if(car != guess1) {
         host <- opts[-c(car, guess1)]
        } else {
         host <- sample(opts[-c(car, guess1)], 1)
        }

        win1 <- guess1 == car # Win no change

        guess2 <- opts[-c(host, guess1)]

        win2 <- guess2 == car # Win change

        if(switch == 0){
                win <- win1
        } else {
                win <- win2
        }

        return(win)
}

#Win probabilities not changing
Prob <- mean(replicate(S, Game(switch = 0)))
Prob
0.3334

#Win probabilities changing
Prob <- mean(replicate(S, Game(switch = 1)))
Prob
0.6654
```

1.2 Bayesian framework: A brief summary of theory

For two random objects θ and $\mathbf{y}$, the Bayes' rule may be analogously used,[5]

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) \times \pi(\theta)}{p(\mathbf{y})}, \tag{1.3}$$

where $\pi(\theta|\mathbf{y})$ is the posterior density function, $\pi(\theta)$ is the prior density, $p(\mathbf{y}|\theta)$ is the likelihood (statistical model), and

$$p(\mathbf{y}) = \int_{\mathbf{\Theta}} p(\mathbf{y}|\theta)\pi(\theta)d\theta = \mathbb{E}\left[p(\mathbf{y}|\theta)\right] \tag{1.4}$$

is the marginal likelihood or prior predictive. Observe that for this expected value to be meaningful the prior should be a proper density, that is, integrates to one, otherwise, it does not make sense.

Observe that $p(\mathbf{y}|\theta)$ is not a density in θ. In addition, $\pi(\theta)$ does not have to integrate to 1, that is, $\pi(\theta)$ can be an improper density function, $\int_{\mathbf{\Theta}} \pi(\theta)d\theta = \infty$. However, $\pi(\theta|\mathbf{y})$ is a proper density function, that is, $\int_{\mathbf{\Theta}} \pi(\theta|\mathbf{y})d\theta = 1$. For instance, set $\pi(\theta) = c$, where c is a constant, then $\int_{\mathbf{\Theta}} cd\theta = \infty$. However, $\int_{\mathbf{\Theta}} \pi(\theta|\mathbf{y})d\theta = \int_{\mathbf{\Theta}} \frac{p(\mathbf{y}|\theta)\times c}{\int_{\mathbf{\Theta}} p(\mathbf{y}|\theta)\times cd\theta}d\theta = 1$ where c cancels out.

$\pi(\theta|\mathbf{y})$ is a sample updated "probabilistic belief" version of $\pi(\theta)$, where $\pi(\theta)$ is a prior probabilistic belief which can be constructed from previous empirical work, theory foundations, expert knowledge and/or mathematical convenience. This prior usually depends on parameters, which are named *hyperparameters*. In addition, the Bayesian approach implies using a probabilistic model about $\mathbf{y}$ given θ, that is, $p(\mathbf{y}|\theta)$, where its integral over $\mathbf{\Theta}$, $p(\mathbf{y})$ is named *the model evidence* due to being a measure of model fit to the data.

Observe that the Bayesian inferential approach is conditional, that is, what can we learn about an unknown object θ given that we already observed $\mathbf{y}$? The answer is also conditional on the probabilistic model, that is $p(\mathbf{y}|\theta)$. So, what if we want to compare different models, let's say $\mathcal{M}_m$, $m = \{1, 2, \ldots, M\}$. Then, we should make explicit this in the Bayes' rule formulation,

$$\pi(\theta|\mathbf{y}, \mathcal{M}_m) = \frac{p(\mathbf{y}|\theta, \mathcal{M}_m) \times \pi(\theta|\mathcal{M}_m)}{p(\mathbf{y}|\mathcal{M}_m)}. \tag{1.5}$$

The posterior model probability is

$$\pi(\mathcal{M}_m|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_m) \times \pi(\mathcal{M}_m)}{p(\mathbf{y})}, \tag{1.6}$$

[5]From a Bayesian perspective θ is fixed, but unknown. Then, it is treated as a random object.

where $p(\mathbf{y}|\mathcal{M}_m) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\theta, \mathcal{M}_m) \times \pi(\theta|\mathcal{M}_m) d\theta$ due to equation 1.5, and $\pi(\mathcal{M}_m)$ is the prior model probability.

Calculating $p(\mathbf{y})$ in equations 1.3 and 1.6 is very demanding most of the realistic cases. Fortunately, it is not required when performing inference about θ as this is integrated out from it. Then, all what you need to know about the shape of θ is in $p(\mathbf{y}|\theta, \mathcal{M}_m) \times \pi(\theta|\mathcal{M}_m)$ or without explicitly conditioning on $\mathcal{M}_m$,

$$\pi(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta) \times \pi(\theta). \tag{1.7}$$

Equation 1.7 is a very good shortcut to perform Bayesian inference about θ.

We also can avoid calculating $p(\mathbf{y})$ when performing model selection (hypothesis testing) using posterior odds ratio, that is, comparing models $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$\begin{aligned} PO_{12} &= \frac{\pi(\mathcal{M}_1|\mathbf{y})}{\pi(\mathcal{M}_2|\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\mathcal{M}_1)}{p(\mathbf{y}|\mathcal{M}_2)} \times \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}, \end{aligned} \tag{1.8}$$

where the first term in equation 1.8 is named the Bayes factor, and the second term is the prior odds. Observe that the Bayes factor is a ratio of ordinates for $\mathbf{y}$ under different models. Then, the Bayes factor is a measure of relative sample evidence in favor of model 1 compared to model 2.

However, we still need to calculate $p(\mathbf{y}|\mathcal{M}_m) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\theta, \mathcal{M}_m)\pi(\theta|\mathcal{M}_m)d\theta = \mathbb{E}\left[p(\mathbf{y}|\theta, \mathcal{M}_m)\right]$. For this integral to be meaningful, the prior must be proper. Using improper prior has unintended consequences when comparing models, for instance, parsimonious models are favored by posterior odds or Bayes factors depend on units of measure.

A nice feature of comparing models using posterior odds is that if we have an exhaustive set of competing models such that $\sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y}) = 1$, then we can recover $\pi(\mathcal{M}_m|\mathbf{y})$ without calculating $p(\mathbf{y})$. In particular, given two models $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $\pi(\mathcal{M}_1|\mathbf{y}) + \pi(\mathcal{M}_2|\mathbf{y}) = 1$. Then, $\pi(\mathcal{M}_1|\mathbf{y}) = \frac{PO_{12}}{1+PO_{12}}$ and $\pi(\mathcal{M}_2|\mathbf{y}) = 1 - \pi(\mathcal{M}_1|\mathbf{y})$. In general, $\pi(\mathcal{M}_m|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_m)\times\pi(\mathcal{M}_m)}{\sum_{l=1}^{M} p(\mathbf{y}|\mathcal{M}_l)\times\pi(\mathcal{M}_l)}$.

Table 1.1 shows guidelines for the interpretation of $2\log(PO_{12})$ [38]. This is done to replicate the structure of the likelihood ratio test statistic. However, posterior odds do not require nested models as the likelihood ratio test does.

Observe that the posterior odds ratio is a relative criterion, that is, we specify an exhaustive set of competing models, and compare them. However, we may want to check the performance of a model in its own or use a non-

TABLE 1.1
Kass and Raftery guidelines.

$2 \times \log(PO_{12})$	PO_{12}	**Evidence against $\mathcal{M}_2$**
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

informative prior. In this case, we can use *the posterior predictive p-value* [25, 26].[6]

The intuition behind the predictive p-value is simple: analyze discrepancy between model's assumptions and data by checking a potential extreme tail-area probability. Observe that this approach does not check if a model is true, its focus is on potential discrepancies between a model and the data at hand.

This is done simulating pseudo-data from our sampling model $(\mathbf{y}^{(s)}, s = 1, 2, \ldots, S)$ using draws from the posterior distribution, and then calculating a discrepancy measure, $D(\mathbf{y}^{(s)}, \theta)$, to estimate the posterior predictive p-value, $p_D(\mathbf{y}) = P[D(\mathbf{y}^{(s)}, \theta) \geq D(\mathbf{y}, \theta)]$ using the proportion of the S draws for which $D(\mathbf{y}^{(s)}, \theta^{(s)}) \geq D(\mathbf{y}, \theta^{(s)})$. Extreme tail probabilities ($p(D_{\mathbf{y}}) \leq 0.05$ or $p(D_{\mathbf{y}}) \geq 0.95$) suggest potential discrepancy between the data and the model. [26] also suggest the posterior predictive p-value based on the *minimum discrepancy*, $D_{min}(\mathbf{y}) = \min_\theta D(\mathbf{y}, \theta)$, and the *average discrepancy* statistic $D(\mathbf{y}) = \mathbb{E}[D(\mathbf{y}, \theta)] = \int_{\boldsymbol{\Theta}} D(\mathbf{y}, \theta)\pi(\theta|\mathbf{y})d\theta$. These alternatives can be more computational demanding.

The Bayesian approach is also suitable to get probabilistic predictions, that is, we can obtain a posterior predictive density

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}, \mathcal{M}_m) &= \int_{\boldsymbol{\Theta}} \pi(\mathbf{Y}_0, \theta|\mathbf{y}, \mathcal{M}_m)d\theta \\ &= \int_{\boldsymbol{\Theta}} \pi(\mathbf{Y}_0|\theta, \mathbf{y}, \mathcal{M}_m)\pi(\theta|\mathbf{y}, \mathcal{M}_m)d\theta. \end{aligned} \tag{1.9}$$

Observe that equation 1.9 is again an expectation $\mathbb{E}[\pi(\mathbf{Y}_0|\theta, \mathbf{y}, \mathcal{M}_m)]$, this time using the posterior distribution. Therefore, the Bayesian approach takes estimation error into account when performing prediction.

As we have shown many times, expectation (integration) is a common feature in Bayesian inference. That is why the remarkable relevance of computation based on *Monte Carlo integration* in the Bayesian framework.

Bayesian model average (BMA) allows considering model uncertainty in prediction or any unknown probabilistic object. In the prediction case,

[6] See also [1] to show potential flows due to using data twice in the construction of the predictive p values, and alternative proposals, for instance *the partial posterior predictive p value*.

$$\pi(\mathbf{Y}_0|\mathbf{y}) = \sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y})\pi(\mathbf{Y}_0|\mathbf{y}, \mathcal{M}_m), \tag{1.10}$$

and parameters case,

$$\pi(\theta|\mathbf{y}) = \sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y})\pi(\theta|\mathbf{y}, \mathcal{M}_m), \tag{1.11}$$

where

$$\mathbb{E}(\theta|\mathbf{y}) = \sum_{m=1}^{M} \hat{\theta}_m \pi(\mathcal{M}_m|\mathbf{y}), \tag{1.12}$$

and

$$Var(\theta|\mathbf{y}) = \sum_{m=1}^{M} \pi(\mathcal{M}_m|\mathbf{y})\widehat{Var}(\theta|\mathbf{y}, \mathcal{M}_m) + \sum_{m=1}^{M} \pi(M_m|\mathbf{y})(\hat{\theta}_m - \mathbb{E}[\theta|\mathbf{y}])^2, \tag{1.13}$$

$\hat{\theta}_m$ and $\widehat{Var}(\theta|\mathbf{y}, \mathcal{M}_m)$ are the posterior mean and variance under model m, respectively.

Observe how the variance in equation 1.13 encloses extra variability due to potential differences between mean posterior estimates associated with each model, and the posterior mean involving model uncertainty in equation 1.12.

A nice advantage of the Bayesian approach, which is very useful in *state space representations* (see Chapter **??**), is the way that the posterior distribution updates with new sample information. Given $\mathbf{y} = \mathbf{y}_{1:t+1}$ a sequence of observations, then

$$\begin{aligned} \pi(\theta|\mathbf{y}_{1:t+1}) &\propto p(\mathbf{y}_{1:t+1}|\theta) \times \pi(\theta) \\ &= p(y_{t+1}|\mathbf{y}_{1:t}, \theta) \times p(\mathbf{y}_{1:t}|\theta) \times \pi(\theta) \\ &\propto p(y_{t+1}|\mathbf{y}_{1:t}, \theta) \times \pi(\theta|\mathbf{y}_{1:t}). \end{aligned} \tag{1.14}$$

We observe that the new prior is just the posterior distribution using the previous observation. This is particular useful under the assumption of *conditional independence*, that is, $y_{t+1} \perp \mathbf{y}_{1:t}|\theta$, then $p(y_{t+1}|\mathbf{y}_{1:t}, \theta) = p(y_{t+1}|\theta)$ such that the posterior can be recovered recursively [51]. This facilities online updating due to all information up to t being in θ. Then, $\pi(\theta|\mathbf{y}_{1:t+1}) \propto p(y_{t+1}|\theta) \times \pi(\theta|\mathbf{y}_{1:t}) \propto \prod_{h=1}^{t+1} p(y_h|\theta) \times \pi(\theta)$. This recursive expression can be

calculated faster at some specific point in time t compared to a batch mode algorithm, which requires processing simultaneously all information up to t.

It is also important to wonder about the sampling properties of "Bayesian estimators". This topic has attracted attention of statisticians and econometricians long time ago. For instance, asymptotic posterior concentration at the population parameter vector is discussed by [10]. Convergence of posterior distributions is stated by the Bernstein-von Mises theorem [42], which creates a link between *credible intervals (sets)* and confidence intervals (sets), where a credible interval is an interval in the domain of the posterior distribution within which an unknown parameter falls with a particular probability. Credible intervals treat bounds as fixed and parameters as random, whereas confidence intervals reverse this. There are many settings in parametric models where Bayesian credible intervals with α level convergences asymptotically to confidence intervals at α level. This suggests that Bayesian inference is asymptotically correct from a sampling perspective in these settings.

A heuristic approach to show this in the simplest case where we assume random sampling and $\theta \in \mathcal{R}$ is the following: $p(\mathbf{y}|\theta) = \prod_{i=1}^{N} p(y_i|\theta)$ such that the log likelihood is $l(\mathbf{y}|\theta) \equiv \log p(\mathbf{y}|\theta) = \sum_{i=1}^{N} \log p(y_i|\theta) = N \times \bar{l}(\mathbf{y}|\theta)$ where $\bar{l} \equiv \frac{1}{N} \sum_{i=1}^{N} \log p(y_i|\theta)$ is the mean likelihood.[7] Then, the posterior distribution is proportional to

$$\begin{aligned} \pi(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta) \times \pi(\theta) \\ &= \exp\left\{N \times \bar{l}(\mathbf{y}|\theta)\right\} \times \pi(\theta). \end{aligned} \tag{1.15}$$

Observe that as the sample size gets large, that is, $N \to \infty$, the exponential term should dominate the prior distribution as long as this does not depend on N such that the likelihood determines the posterior distribution asymptotically.

Maximum likelihood theory shows that $\lim_{N\to\infty} \bar{l}(\mathbf{y}|\theta) \to \bar{l}(\mathbf{y}|\theta_0)$ where θ_0 is the population parameter of the data generating process. In addition, doing a second order Taylor expansion of the log likelihood at the Maximum likelihood estimator,

$$\begin{aligned} l(\mathbf{y}|\theta) &\approx l(\mathbf{y}|\hat{\theta}) + \left.\frac{dl(\mathbf{y}|\theta)}{d\theta}\right|_{\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2} \left.\frac{d^2 l(\mathbf{y}|\theta)}{d\theta^2}\right|_{\hat{\theta}} (\theta - \hat{\theta})^2 \\ &= l(\mathbf{y}|\hat{\theta}) + \frac{1}{2} \sum_{i=1}^{N} \left.\frac{d^2 l(y_i|\theta)}{d\theta^2}\right|_{\hat{\theta}} (\theta - \hat{\theta})^2 \\ &= l(\mathbf{y}|\hat{\theta}) - \frac{1}{2} N \left[-\bar{l}''\big|_{\hat{\theta}}\right] (\theta - \hat{\theta})^2 \end{aligned}$$

[7] Take into account that in the likelihood function the argument is θ. However, we keep the notation for facility in exposition.

$$= l(\mathbf{y}|\hat{\theta}) - \frac{N}{2\sigma^2}(\theta - \hat{\theta})^2$$

where $\frac{dl(\mathbf{y}|\theta)}{d\theta}\Big|_{\hat{\theta}} = 0$, $\bar{l}'' \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{d^2 l(y_i|\theta)}{d\theta^2}\Big|_{\hat{\theta}}$ and $\sigma^2 := \left[-\bar{l}''\big|_{\hat{\theta}}\right]^{-1}$. Then,

$$\begin{aligned}
\pi(\theta|\mathbf{y}) &\propto \exp\{l(\mathbf{y}|\theta)\} \times \pi(\theta) \\
&\approx \exp\left\{l(\mathbf{y}|\hat{\theta}) - \frac{N}{2\sigma^2}(\theta - \hat{\theta})^2\right\} \times \pi(\theta) \\
&\propto \exp\left\{-\frac{N}{2\sigma^2}(\theta - \hat{\theta})^2\right\} \times \pi(\theta)
\end{aligned}$$

Observe that we have that the posterior density is proportional to the kernel of a normal density with mean $\hat{\theta}$ and variance σ^2/N as long as $\pi(\hat{\theta}) \neq 0$. This kernel dominates as the sample size gets large due to N in the exponential term. Observe that the prior should not exclude values of θ that are logically possible, such as $\hat{\theta}$.

1.2.1 Example: Health insurance

Suppose that you are analyzing to buy a health insurance next year. To make a better decision you want to know *what is the probability that you visit your Doctor at least once next year?* To answer this question you have records of the number of times that you have visited your Doctor the last 5 years, $\mathbf{y} = \{0, 3, 2, 1, 0\}$. How to proceed?

Assuming that this is a random sample[8] from a data generating process (statistical model) that is Poisson, that is, $Y_i \sim P(\lambda)$, and your probabilistic prior beliefs about λ are well described by a Gamma distribution with shape and scale parameters α_0 and β_0, $\lambda \sim G(\alpha_0, \beta_0)$, then, you are interested in calculating the probability $P(Y_0 > 0|\mathbf{y})$. You need to calculate the posterior predictive density $\pi(Y_0|\mathbf{y})$ to answer this question in a Bayesian way.

In this example, $p(\mathbf{y}|\lambda)$ is Poisson, and $\pi(\lambda)$ is Gamma. Then, using 1.9

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

where the posterior distribution is $\pi(\lambda|\mathbf{y}) \propto \lambda^{\sum_{i=1}^{N} y_i + \alpha_0 - 1}\exp\left\{-\lambda\left(\frac{\beta_0 N + 1}{\beta_0}\right)\right\}$ by equation 1.3. $\Gamma(\cdot)$ is the gamma function.

[8] Independent and identically distributed draws.

Observe that the last expression is the kernel of a Gamma distribution with parameters $\alpha_n = \sum_{i=1}^{N} y_i + \alpha_0$ and $\beta_n = \frac{\beta_0}{\beta_0 N+1}$. Given that $\int_0^\infty \pi(\lambda|\mathbf{y})d\lambda = 1$, then the constant of proportionality in the last expression is $\Gamma(\alpha_n)\beta_n^{\alpha_n}$. The posterior density function $\pi(\lambda|\mathbf{y})$ is $G(\alpha_n, \beta_n)$.

Observe that

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

where $\bar{y}$ is the sample mean, which is the maximum likelihood estimator of λ, $w = \left(\frac{N\beta_0}{N\beta_0+1}\right)$ and $\mathbb{E}[\lambda] = \alpha_0\beta_0$ is the prior mean. The posterior mean is a weighted average of the maximum likelihood estimator (sample information) and the prior mean. Observe that $\lim_{N\to\infty} w = 1$, that is, the sample information asymptotically dominates.

The predictive distribution is

$$\begin{aligned}\pi(Y_0|\mathbf{y}) =& \int_0^\infty \frac{\lambda^{y_0}\exp\{-\lambda\}}{y_0!} \times \frac{1}{\Gamma(\alpha_n)\beta_n^{\alpha_n}}\lambda^{\alpha_n-1}\exp\{-\lambda/\beta_n\}\,d\lambda \\ =& \frac{1}{y_0!\Gamma(\alpha_n)\beta_n^{\alpha_n}}\int_0^\infty \lambda^{y_0+\alpha_n-1}\exp\left\{-\lambda\left(\frac{1+\beta_n}{\beta_n}\right)\right\}d\lambda \\ =& \frac{\Gamma(y_0+\alpha_n)\left(\frac{\beta_n}{\beta_n+1}\right)^{y_0+\alpha_n}}{y_0!\Gamma(\alpha_n)\beta_n^{\alpha_n}} \\ =& \binom{y_0+\alpha_n-1}{y_0}\left(\frac{\beta_n}{\beta_n+1}\right)^{y_0}\left(\frac{1}{\beta_n+1}\right)^{\alpha_n}.\end{aligned}$$

The third equality follows from the kernel of a Gamma density, and the fourth from $\binom{y_0+\alpha_n-1}{y_0} = \frac{(y_0+\alpha_n-1)(y_0+\alpha_n-2)\ldots\alpha_n}{y_0!} = \frac{\Gamma(y_0+\alpha_n)}{\Gamma(\alpha_n)y_0!}$ using a property of the Gamma function.

Observe that this is a Negative Binomial density, that is $Y_0|\mathbf{y} \sim NB(\alpha_n, p_n)$ where $p_n = \frac{\beta_n}{\beta_n+1}$.

A key question is how to fix the hyperparameters. In this exercise we use two approaches for exposition purposes. We set $\alpha_0 = 0.001$ and $\beta_0 = 1/0.001$ which imply vague prior information about λ due to having a large degree of variability compared to the mean information.[9] In particular, $\mathbb{E}[\lambda] = 1$ and $\mathbb{V}ar[\lambda] = 1000$.

[9] We should be aware that there may be technical problems using this king of hyperparameters in this setting [27].

In this setting, $P(Y_0 > 0|\mathbf{y}) = 1 - P(Y_0 = 0|\mathbf{y}) \approx 0.67$. That is, the probability of visiting the Doctor at least once next year is approximately 0.67.

Another approach is using *Empirical Bayes*, where we set the hyper-parameters maximizing the logarithm of the marginal likelihood, that is, $\left[\hat{\alpha}_0 \ \hat{\beta}_0\right]^\top = \underset{\alpha_0,\beta_0}{\text{argmax}} \ \ln p(\mathbf{y})$ where

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

Using the empirical Bayes approach, we get $\hat{\alpha}_0 = 51.8$ and $\hat{\beta}_0 = 0.023$, then $P(Y_0 > 0|\mathbf{y}) = 1 - P(Y_0 = 0|\mathbf{y}) \approx 0.70$.

Observe that we can calculate the posterior odds comparing the model using an Empirical Bayes prior (model 1) versus the vague prior (model 2). We assume that $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2) = 0.5$, then

$$\begin{aligned} PO_{12} &= \frac{p(\mathbf{y}|\text{Empirical Bayes})}{p(\mathbf{y}|\text{Vague prior})} \\ &= \frac{\frac{\Gamma(\sum_{i=1}^N y_i + 51.807)\left(\frac{0.023}{N\times 0.023+1}\right)^{\sum_{i=1}^N y_i}\left(\frac{1}{N\times 0.023+1}\right)^{51.807}}{\Gamma(51.807)}}{\frac{\Gamma(\sum_{i=1}^N y_i + 0.001)\left(\frac{1/0.001}{N/0.001+1}\right)^{\sum_{i=1}^N y_i}\left(\frac{1}{N/0.001+1}\right)^{0.001}}{\Gamma(0.001)}} \\ &\approx 919. \end{aligned}$$

Then, $2 \times \log(PO_{12}) = 13.64$, there is very strong evidence against the vague prior model (see Table 1.1). In particular, $\pi(\text{Empirical Bayes}|\mathbf{y}) = \frac{919}{1+919} = 0.999$ and $\pi(\text{Vague prior}|\mathbf{y}) = 1 - 0.999 = 0.001$. These probabilities can be used to perform Bayesian model average (BMA). In particular,

$$\begin{aligned} \mathbb{E}(\lambda|\mathbf{y}) &= 1.2 \times 0.999 + 1.2 \times 0.001 = 1.2 \\ Var(\lambda|\mathbf{y}) &= 0.025 \times 0.999 + 0.24 \times 0.001 \\ &\quad + (1.2 - 1.2)^2 \times 0.999 + (1.2 - 1.2)^2 \times 0.001 = 0.025. \end{aligned}$$

The BMA predictive distribution is a mix of negative binomial distributions, that is, $y_0|\mathbf{y} \sim 0.999 \times NB(57.8, 0.02) + 0.001 \times NB(6.001, 0.17)$.

R code. Health insurance, predictive distribution using vague hyperparameters

```
set.seed(010101)
y <- c(0, 3, 2, 1, 0) # Data
N <- length(y)

ProbBo <- function(y, a0, b0){
        N <- length(y)
        #sample size
        an <- a0 + sum(y)
        # Posterior shape parameter
        bn <- b0 / ((b0 * N) + 1)
        # Posterior scale parameter
        p <- bn / (bn + 1)
        # Probability negative binomial density
        Pr <- 1 - pnbinom(0, size=an,prob=(1 - p))
        # Probability of visiting the Doctor
        # at least once next year
        # Observe that in R there is a slightly
        # different parametrization.
        return(Pr)
}

# Using a vague prior:

a0 <- 0.001 # Prior shape parameter
b0 <- 1 / 0.001 # Prior scale parameter
PriMeanV <- a0 * b0 # Prior mean
PriVarV <- a0 * b0^2 # Prior variance
Pp <- ProbBo(y, a0 = 0.001, b0 = 1 / 0.001)
# This setting is vague prior information.
Pp
0.67
```

R code. Health insurance, predictive distribution using empirical Bayes

```
# Using Emprirical Bayes
LogMgLik <- function(theta, y){
 N <- length(y)
 #sample size
 a0 <- theta[1]
 # prior shape hyperparameter
 b0 <- theta[2]
 # prior scale hyperparameter
 an <- sum(y) + a0
 # posterior shape parameter
 if(a0 <= 0 || b0 <= 0){
  #Avoiding negative values
  lnp <- -Inf
  }else{
  lnp <- lgamma(an) + sum(y)*log(b0/(N*b0+1)) -
   a0*log(N*b0+1) - lgamma(a0)
 }
 # log marginal likelihood
 return(-lnp)
}

theta0 <- c(0.01, 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$convergence
0
a0EB <- EmpBay$par[1]
# Prior shape using empirical Bayes
a0EB
51.81
b0EB <- EmpBay$par[2]
# Prior scale using empirical Bayes
b0EB
0.023
PriMeanEB <- a0EB * b0EB
# Prior mean
PriVarEB <- a0EB * b0EB^2
# Prior variance
PpEB <- ProbBo(y, a0 = a0EB, b0 = b0EB)
# This setting is using emprical Bayes.
PpEB
0.70
```

R code. Health insurance, density plots

```
# Density figures:
# This code helps plotting densities
lambda <- seq(0.01, 10, 0.01)
# Values of lambda
VaguePrior <- dgamma(lambda,shape=a0,scale = b0)
EBPrior <- dgamma(lambda,shape=a0EB,scale = b0EB)
PosteriorV <- dgamma(lambda, shape = a0 + sum(y),
scale = b0 / ((b0 * N) + 1))
PosteriorEB <- dgamma(lambda, shape = a0EB+sum(y),
scale = b0EB / ((b0EB * N) + 1))

# Likelihood function
Likelihood <- function(theta, y){
 LogL <- dpois(y, theta, log = TRUE)
 Lik <- prod(exp(LogL))
 return(Lik)
}
Liks <- sapply(lambda, function(par) {
 Likelihood(par, y = y)})
Sc <- max(PosteriorEB)/max(Liks)
#Scale for displaying in figure
LiksScale <- Liks * Sc
```

Figure 1.2 displays prior and posterior densities based on vague and Empirical Bayes hyperparameters. We see that prior and posterior densities using the latter are more informative as expected.

Figure 1.3 shows the prior, scaled likelihood and posterior densities of λ based on the hyperparameters of the Empirical Bayes approach. The posterior density a compromise between prior and sample information.

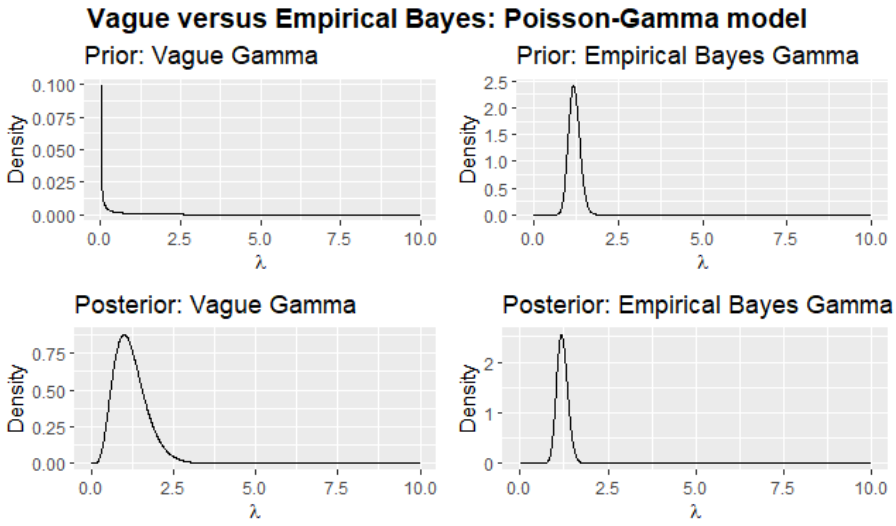

FIGURE 1.2
Vague versus Empirical Bayes: Poisson-Gamma model.

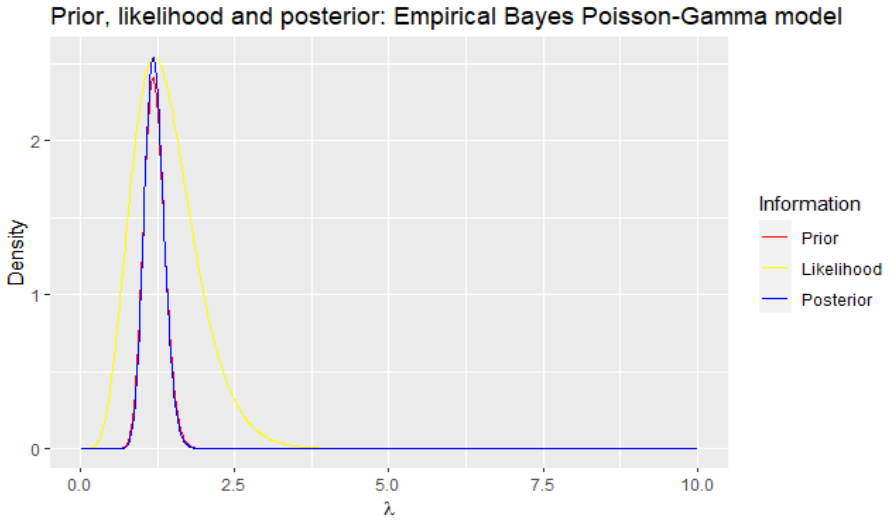

FIGURE 1.3
Prior, likelihood and posterior: Empirical Bayes Poisson-Gamma model.

R code. Health insurance, Predictive density

```
# Predictive distributions
PredDen <- function(y, y0, a0, b0){
        N <- length(y)
        #sample size
        an <- a0 + sum(y)
        # Posterior shape parameter
        bn <- b0 / ((b0 * N) + 1)
        # Posterior scale parameter
        p <- bn / (bn + 1)
        # Probability negative binomial density
        Pr <- dnbinom(y0, size=an, prob=(1 - p))
        # Predictive density
        # Observe that in R there is a slightly
        # different parametrization.
        return(Pr)
}
y0 <- 0:10
PredVague <- PredDen(y=y, y0=y0, a0=a0, b0=b0)
PredEB <- PredDen(y=y, y0=y0, a0=a0EB, b0=b0EB)
```

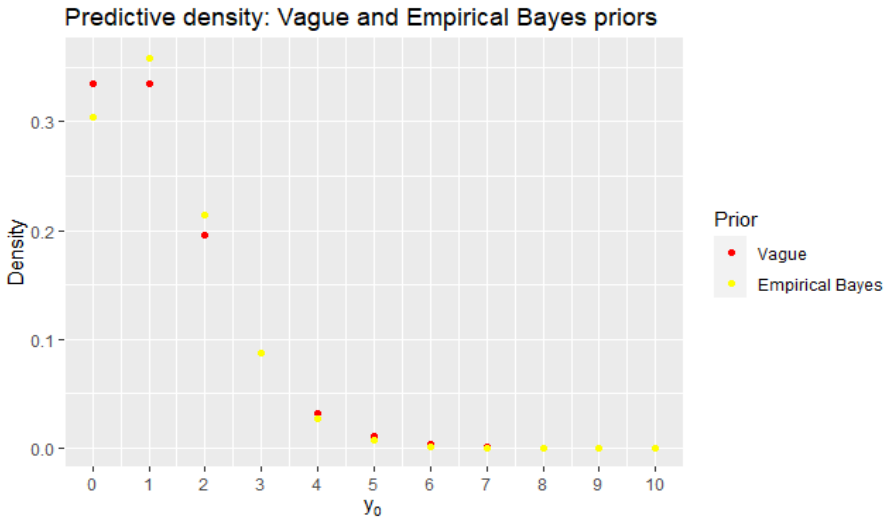

FIGURE 1.4
Predictive density: Vague and Empirical Bayes.

Figure 1.4 displays the predictive probability mass of not having any physi-

cian visit next year, having one, two, and so on using Empirical Bayes and vague hyperparameters. The predictive probability of no having any visit are approximately equal to 30% and 33% based on the Empirical Bayes and vague hyperparameters.

R code. Health insurance, Bayesian model average

```
# Posterior odds: Vague vs Empirical Bayes

PO12 <- exp(-LogMgLik(c(a0EB, b0EB), y = y))
                /exp(-LogMgLik(c(a0, b0), y = y))

PO12
919

PostProMEM <- PO12/(1 + PO12)
PostProMEM
0.998
# Posterior model probability Empirical Bayes
PostProbMV <- 1 - PostProMEM
PostProbMV
0.002
# Posterior model probability vague prior

# Bayesian model average (BMA)
PostMeanEB <- (a0EB + sum(y)) * (b0EB / (b0EB * N + 1))
# Posterior mean Empirical Bayes
PostMeanV <- (a0 + sum(y)) * (b0 / (b0 * N + 1))
# Posterior mean vague priors
BMAmean <- PostProMEM * PostMeanEB
                + PostProbMV * PostMeanV
BMAmean
1.2
# BMA posterior mean

PostVarEB <- (a0EB + sum(y)) * (b0EB/(b0EB * N + 1))^2
# Posterior variance Empirical Bayes
PostVarV <- (a0 + sum(y)) * (b0 / (b0 * N + 1))^2
# Posterior variance vague prior

BMAVar <- PostProMEM * PostVarEB + PostProbMV*PostVarV
                + PostProMEM * (PostMeanEB - BMAmean)^2
                + PostProbMV * (PostMeanV - BMAmean)^2
# BMA posterior variance
BMAVar
0.025

# BMA: Predictive
BMAPred <- PostProMEM * PredEB+PostProbMV * PredVague
```

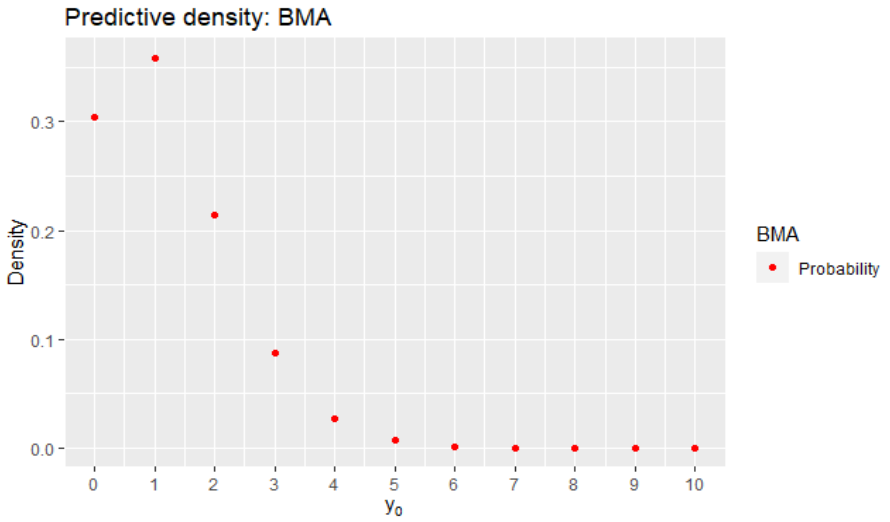

FIGURE 1.5
Bayesian model average: Predictive density.

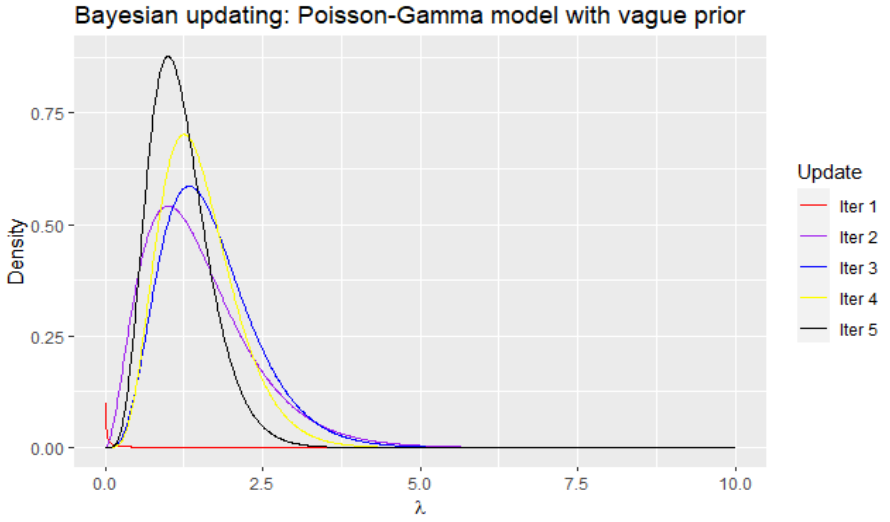

FIGURE 1.6
Bayesian updating: Posterior densities.

Figure 1.5 displays the predictive density based on the vague and Empirical Bayes hyperparameters. This figures resembles basically the predictive based

on the Empirical Bayes framework as the posterior model probability of that setting is almost one.

Figure 1.6 displays how the posterior distribution updates given new sample information based on an initial non-informative prior (iteration 1). We see that iteration 5 is based on all the sample information in our example, as a consequence, the posterior density in iteration 5 is equal to the posterior density in Figure 1.3.

R code. Health insurance, Bayes updating

```
# Bayesian updating
BayUp <- function(y, lambda, a0, b0){
        N <- length(y)
        #sample size
        an <- a0 + sum(y)
        # Posterior shape parameter
        bn <- b0 / ((b0 * N) + 1)
        # Posterior scale parameter
        p <- dgamma(lambda, shape = an, scale = 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 = 0.001, b0 = 1/0.001)}
        else{
                PostUpi <- BayUp(y[i], lambda,
                a0 = PostUpi$a0New, b0 = PostUpi$b0New)
        }
        PostUp <- cbind(PostUp, PostUpi$Post)
}
```

1.3 Bayesian reports: Decision theory under uncertainty

The Bayesian framework allows reporting the full posterior distributions. However, some situations demand to report a specific value of the posterior distribution (point estimate), an informative interval (set), point or interval predictions and/or selecting a specific model. Decision theory offers an elegant

framework to make a decision regarding what are the optimal posterior values to report [8].

The point of departure is a *loss function*, which is a non-negative real value function whose arguments are the unknown *state of nature* ($\mathbf{\Theta}$), and a set of *actions* to be made ($\mathcal{A}$), that is,

$$L(\theta, a) : \mathbf{\Theta} \times \mathcal{A} \to \mathcal{R}^{+}.$$

This function is a mathematical expression of the loss of making mistakes. In particular, selecting action $a \in \mathcal{A}$ when $\theta \in \mathbf{\Theta}$ is the true. In our case, the unknown state of nature can be parameters, functions of them, future or unknown realizations, models, etc.

From a Bayesian perspective, we should choose the action that minimizes the posterior expected loss ($a^{*}(\mathbf{y})$), that is the *posterior risk function* ($\mathbb{E}[L(\theta, a)|\mathbf{y}]$),

$$a^{*}(\mathbf{y}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\ \mathbb{E}[L(\theta, a)|\mathbf{y}],$$

where $\mathbb{E}[L(\theta, a)|\mathbf{y}] = \int_{\mathbf{\Theta}} L(\theta, a)\pi(\theta|\mathbf{y})d\theta$.[10]

Different loss functions imply different optimal decisions. We illustrate this assuming $\theta \in \mathcal{R}$.

- The quadratic loss function, $L(\theta, a) = [\theta - a]^2$, gives as optimal decision the posterior mean, $a^{*}(\mathbf{y}) = \mathbb{E}[\theta|\mathbf{y}]$, that is

  $$\mathbb{E}[\theta|\mathbf{y}] = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \int_{\Theta} [\theta - a]^2 \pi(\theta|\mathbf{y})d\theta.$$

  To get this results, let us use the first condition order, differentiate the risk function with respect to a, interchange differential and integral order, and set this equal to zero, $-2\int_{\Theta}[\theta - a^{*}]\pi(\theta|\mathbf{y})d\theta = 0$ implies that $a^{*}\int_{\Theta}\pi(\theta|\mathbf{y})d\theta = a^{*}(\mathbf{y}) = \int_{\Theta}\theta\pi(\theta|\mathbf{y})d\theta = \mathbb{E}[\theta|\mathbf{y}]$, that is, the posterior mean is the Bayesian optimal action. This means that we should report the posterior mean as a point estimate of θ when facing the quadratic loss function.

- The generalized quadratic loss function, $L(\theta, a) = w(\theta)[\theta - a]^2$, where $w(\theta) > 0$ is a weighting function, gives as optimal decision rule the weighted mean. We should follow same steps as the previous result to get $a^{*}(\mathbf{y}) = \frac{\mathbb{E}[w(\theta)\times\theta|\mathbf{y}]}{\mathbb{E}[w(\theta)|\mathbf{y}]}$. Observe that the weighted average is driven by the weighted function $w(\theta)$.

[10] [13] propose Laplace type estimators (LTE) based on the *quasi-posterior*, $p(\theta) = \frac{\exp\{L_n(\theta)\}\pi(\theta)}{\int_{\mathbf{\Theta}}\exp\{L_n(\theta)\}\pi(\theta)d\theta}$ where $L_n(\theta)$ is not necessarily a log-likelihood function. The LTE minimizes the *quasi-posterior risk*.

- The absolute error loss function, $L(\theta, a) = |\theta - a|$, gives as optimal action the posterior median (exercise 5).
- The generalized absolute error function,

$$L(\theta, a) = \begin{Bmatrix} K_0(\theta - a), \theta - a \geq 0 \\ K_1(a - \theta), \theta - a < 0 \end{Bmatrix}, \ K_0, K_1 > 0,$$

implies the following risk function,

$$\mathbb{E}[L(\theta, a)|\mathbf{y}] = \int_{-\infty}^{a} K_1(a - \theta)\pi(\theta|\mathbf{y})d\theta + \int_{a}^{\infty} K_0(\theta - a)\pi(\theta|\mathbf{y})d\theta.$$

Differentiating with respect to a, interchanging differentials and integrals, and equating to zero,

$$K_1 \int_{-\infty}^{a^*} \pi(\theta|\mathbf{y})d\theta - K_0 \int_{a^*}^{\infty} \pi(\theta|\mathbf{y})d\theta = 0,$$

then, $\int_{-\infty}^{a^*} \pi(\theta|\mathbf{y})d\theta = \frac{K_0}{K_0+K_1}$, that is, any $K_0/(K_0 + K_1)$-percentile of $\pi(\theta|\mathbf{y})$ is an optimal Bayesian estimate of θ.

We can also use decision theory under uncertainty in hypothesis testing. In particular, testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_0 \cup \Theta_1$ and $\emptyset = \Theta_0 \cap \Theta_1$, there are two actions of interest, a_0 and a_1, where a_j denotes no rejecting H_j, $j = \{0, 1\}$.

Given the $0 - K_j$ loss function,

$$L(\theta, a_j) = \begin{Bmatrix} 0, & \theta \in \Theta_j \\ K_j, & \theta \in \Theta_j, j \neq i \end{Bmatrix},$$

where there is no loss if the right decision is made, for instance, no rejecting H_0 when $\theta \in \Theta_0$, and the loss is K_j when an error is made, for instance, type I error, rejecting the null hypothesis (H_0) when it is true ($\theta \in \Theta_0$), implies a loss equal to K_1 due to picking a_1, no rejecting H_1.

The posterior expected loss associated with decision a_j, that is, no rejecting H_j, is $\mathbb{E}[L(\theta, a_j)|\mathbf{y}] = 0 \times P(\Theta_j|\mathbf{y}) + K_j P(\Theta_i|\mathbf{y}) = K_j P(\Theta_i|\mathbf{y})$, $j \neq i$. Therefore, the Bayes optimal decision is the one that gives the smallest posterior expected loss, that is, the null hypothesis is rejected (a_1 is not rejected), when $K_0 P(\Theta_1|\mathbf{y}) > K_1 P(\Theta_0|\mathbf{y})$. Given our framework ($\Theta = \Theta_0 \cup \Theta_1, \emptyset = \Theta_0 \cap \Theta_1$), then $P(\Theta_0|\mathbf{y}) = 1 - P(\Theta_1|\mathbf{y})$, and as a consequence, $P(\Theta_1|\mathbf{y}) > \frac{K_1}{K_1+K_0}$, that is, the rejection region of the Bayesian test is $R = \left\{\mathbf{y} : P(\Theta_1|\mathbf{y}) > \frac{K_1}{K_1+K_0}\right\}$.

Decision theory also helps to construct interval (region) estimates. Let $\Theta_{C(\mathbf{y})} \subset \Theta$ a *credible set* for θ, and $L(\theta, \Theta_{C(\mathbf{y})}) = 1 - \mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\}$, where

$$\mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\} = \begin{Bmatrix} 1, & \theta \in \Theta_{C(\mathbf{y})} \\ 0, & \theta \notin \Theta_{C(\mathbf{y})} \end{Bmatrix}.$$

Then,

$$L(\theta, \Theta_{C(\mathbf{y})}) = \begin{Bmatrix} 0, & \theta \in \Theta_{C(\mathbf{y})} \\ 1, & \theta \notin \Theta_{C(\mathbf{y})} \end{Bmatrix},$$

where the 0–1 loss function is equal to zero if $\theta \in \Theta_{C(\mathbf{y})}$, and one if $\theta \notin \Theta_{C(\mathbf{y})}$. Then, the risk function is $1 - P(\theta \in \Theta_{C(\mathbf{y})})$.

Given a *measure of credibility* ($\alpha(\mathbf{y})$) that defines the level of trust that $\theta \in \Theta_{C(\mathbf{y})}$; then, we can measure the accuracy of the report by $L(\theta, \alpha(\mathbf{y})) = [\mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\} - \alpha(\mathbf{y})]^2$. This loss function could be used to suggest a choice of the report $\alpha(\mathbf{y})$. Given that this is a quadratic loss function, the optimal action is the posterior mean, that is $\mathbb{E}[\mathbb{1}\left\{\theta \in \Theta_{C(\mathbf{y})}\right\}|\mathbf{y}] = P(\theta \in \Theta_{C(\mathbf{y})}|\mathbf{y})$. This probability can be calculated given the posterior distribution, that is, $P(\theta \in \Theta_{C(\mathbf{y})}|\mathbf{y}) = \int_{\Theta_{C(\mathbf{y})}} \pi(\theta|\mathbf{y})d\theta$. This is a measure of the belief that $\theta \in \Theta_{C(\mathbf{y})}$ given the prior beliefs and sample information.

The set $\Theta_{C(\mathbf{y})} \in \Theta$ is a $100(1-\alpha)\%$ credible set with respect to $\pi(\theta|\mathbf{y})$ if $P(\theta \in \Theta_{C(\mathbf{y})}|\mathbf{y}) = \int_{\Theta_{C(\mathbf{y})}} \pi(\theta|\mathbf{y}) = 1 - \alpha$.

Two alternatives to report credible sets are the *symmetric credible set* and the *highest posterior density set* (HPD). The former is based on $\frac{\alpha}{2}\%$ and $(1 - \frac{\alpha}{2})\%$ percentiles of the posterior distribution, and the latter is a $100(1-\alpha)\%$ credible interval for θ with the property that it has the smallest distance compared to any other $100(1-\alpha)\%$ credible interval for θ based on the posterior distribution. That is, $C(\mathbf{y}) = \{\theta : \pi(\theta|\mathbf{y}) \geq k(\alpha)\}$, where $k(\alpha)$ is the largest number such that $\int_{\theta:\pi(\theta|\mathbf{y})\geq k(\alpha)} \pi(\theta|\mathbf{y})d\theta = 1 - \alpha$. The HPDs can be a collection of disjoint intervals when working with multimodal posterior densities. In addition, they have the limitation of not necessary being invariant under transformations.

Decision theory can be used to perform prediction (point, sets or probabilistic). Suppose that there is a loss function $L(Y_0, a)$ involving the prediction of Y_0. Then, $\mathbb{E}_{Y_0}[L(Y_0, a)] = \int_{\mathcal{Y}_0} L(Y_0, a)\pi(Y_0|\mathbf{y})dY_0$, where $\pi(Y_0|\mathbf{y})$ is the predictive density function. Thus, we make an optimal choice for prediction that minimizes the risk function given a specific loss function.

BMA allows incorporating model uncertainty in a regression framework, sometimes it is desirable to select just one model. A compelling alternative is the model with the highest posterior model probability. This model is the best alternative for prediction in the case of a 0–1 loss function [16].

1.3.1 Example: Health insurance continues

We show some optimal rules in the health insurance example. In particular, the best point estimates of λ given the quadratic, absolute and generalized

absolute loss functions. For the latter, we assume that underestimating λ is twice as costly as overestimating it, that is, $K_0 = 2$ and $K_1 = 1$.

Taking into account that the posterior distribution of λ is $G(\alpha_0 + \sum_{i=1}^{N} y_i, \beta_0/(\beta_0 N + 1))$, using the hyperparameters from empirical Bayes, we have that $\mathbb{E}[\lambda|\mathbf{y}] = \alpha_n\beta_n = 1.2$, the median is 1.19, and the 2/3-th quantile is 1.26. Those are the optimal point estimates for the quadratic, absolute and generalized absolute loss functions.

In addition, we test the null hypothesis $H_0.\lambda \in [0, 1)$ versus $H_1.\lambda \in [1, \infty)$ setting $K_0 = K_1 = 1$ we should reject the null hypothesis due to $P(\lambda \in [0, 1)) = 0.9 > K_1/(K_0 + K_1) = 0.5$.

We get that the 95% symmetric credible interval is (0.91, 1.53), and the highest posterior density interval is (0.9, 1.51). Finally, the optimal point prediction under a quadratic loss function is 1.2, which is the mean value of the posterior predictive distribution, and the optimal model assuming a 0-1 loss function is the model using the hyperparameters from the empirical Bayes procedure due to the posterior model probability of this model being approximately 1, whereas the posterior model probability of the model using vague hyperparameters is approximately 0.

R code. Health insurance, Bayesian reports

```
an <- sum(y) + a0EB
# Posterior shape parameter
bn <- b0EB / (N*b0EB + 1)
# Posterior scale parameter
S <- 1000000
# Number of posterior draws
Draws <- rgamma(1000000, shape = an, scale = bn)
# Posterior draws
###### Point estimation ########
OptQua <- an*bn
# Mean: Optimal choice quadratic loss function
OptQua
OptAbs <- qgamma(0.5, shape = an, scale = bn)
# Median: Optimal choice absolute loss function
OptAbs
# Setting K0 = 2 and K1 = 1, that is, to underestimate
lambda is twice as costly as to overestimate it.
K0 <- 2; K1 <- 1
OptGenAbs <- quantile(Draws, K0/(K0 + K1))
# Median: Optimal choice generalized absolute loss
function
OptGenAbs
```

R code. Health insurance, Bayesian reports continue

```
###### Hypothesis test ########
# H0: lambda in [0,1) vs H1: lambda in [1, Inf]
K0 <- 1; K1 <- 1
ProbH0 <- pgamma(1, shape = an, scale = bn)
ProbH0 # Posterior  probability H0
ProbH1 <- 1 -ProbH0
ProbH1 # Posterior  probability H1
# we should reject H0 given ProbH1 > K1 / (K0 + K1)
###### Credible intervals ########
LimInf <- qgamma(0.025, shape = an, scale = bn)
# Lower bound
LimInf
LimSup <- qgamma(0.975, shape = an, scale = bn)
# Upper bound
LimSup
HDI <- HDInterval::hdi(Draws, credMass = 0.95)
# Highest posterior density credible interval
HDI
###### Predictive optimal choices ########
p <- bn / (bn + 1)
# Probability negative binomial density
OptPred <- p/(1-p)*an
# Optimal point prediction given a quadratic loss
function
 in prediction
OptPred

# Given a 0-1 loss function for prediction, the optimal
model is the one using empirical Bayes due to having
a posterior model probability approximately equal to 1.
```

1.4 Summary

In this chapter we introduce the Bayes' rule to update probabilistic statements using funny examples. Then, we study the three probabilistic objects of main relevance in Bayesian inference: the posterior distribution, the marginal likelihood and the predictive density. The first allows performing inference regarding parameters, the second is required to perform hypothesis test for model selection using the Bayes factor, and the third to perform probabilistic

predictions. We also review some sampling properties of Bayesian estimators, and Bayes update. All those concepts were developed using a simple example in R software. Finally, we introduce some concepts of decision theory that can be used to report summary statistics minimizing posterior expected losses.

1.5 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?*

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?

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

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).

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.

2

Conceptual differences: Bayesian and Frequentist approaches

We give some of the conceptual differences between the Bayesian and Frequentist inferential approaches. We emphasize in the Bayesian concepts as most of the readers can be familiarized with the Frequentist statistical framework.

2.1 The concept of probability

Let us begin with the following thought experiment: Assume that you are watching the international game show "Who wants to be a millionaire?", the contestant is asked to answer a very simple question: **What is the last name of the brothers who are credited with inventing the world's first successful motor-operated airplane?**

- What is the probability that the contestant answers this question correctly?

  Unless you have:

  1. watched this particular contestant participating in this show many times,
  2. seen him asked this same question each time,
  3. and computed the relative frequency with which he gives the correct answer,

  you need to answer this question as a Bayesian!

Uncertainty about the event *answer this question* needs to be expressed as a "degree of belief" informed both by information coming from data on the skill of the particular participant, and how much he knows about inventors, and possibly prior knowledge on his performance in other game shows. Of course, your prior knowledge of the contestant may be minimal, or it may be very informed. Either way, your final answer remains a degree of belief held about an uncertain, and inherently unrepeatable state of nature.

The point of this hypothetical, light-hearted scenario is simply to highlight that a key distinction between the Frequentist and Bayesian approaches to inference is not the use (or nature) of prior information, but simply the manner

in which probability is used. To the Bayesian, probability is the mathematical construct used to quantify uncertainty about an unknown state of nature, conditional on observed data and prior knowledge about the context in which that state of nature occurs. To the Frequentist, probability is linked intrinsically to the concept of a repeated experiment, and the relative frequency with which a particular outcome occurs, conditional on that unknown state. This distinction remains key whether the Bayesian chooses to be *informative or subjective* in the specification of prior information, or chooses to be *non-informative or objective*.

Frequentists consider probability as a physical phenomenon, like mass or wavelength, whereas Bayesians stipulate that probability lives in the mind of scientists as any scientific construct [49].

It seems that the understanding of the concept of probability for the common human being is more associated with "degrees of belief" rather than relative frequency. Peter Diggle, President of The Royal Statistical Society (2014-2016), was asked "A different trend which has surged upwards in statistics during Peter's career is the popularity of Bayesian statistics. Does Peter consider himself a Bayesian?", and he replied "... you can't not believe in Bayes' theorem because it's true. But that doesn't make you a Bayesian in the philosophical sense. When people are making personal decisions – even if they don't formally process Bayes' theorem in their mind – they are adapting what they think they should believe in response to new evidence as it comes in. Bayes' theorem is just the formal mathematical machinery for doing that."

However, we should say that psychological experiments suggest that human beings suffer from *anchoring*, that is, a cognitive bias that causes us to rely too heavily on the previous information (prior) such that the updating process (posterior) due to new information (likelihood) being low compared to the Bayes' rule [36].

2.2 Subjectivity is not the key

The concepts of *subjectivity* and *objectivity* indeed characterize both statistical paradigms in differing ways. Among Bayesians there are those who are immersed in *subjective* rationality [53, 18, 54, 43], but others who adopt *objective* prior distributions such as Jeffreys', reference, empirical or robust [3, 40, 34, 7] to operationalize Bayes' rule, and thereby weight quantitative (data-based) evidence. Among Frequentists, there are choices made about significance levels which, if not explicitly subjective, are typically not grounded in any objective and documented assessment of the relative losses of Type I and Type II er-

rors.[1] In addition, both Frequentist and Bayesian statisticians make decisions about the form of the data generating process, or "model", which – if not subject to rigorous diagnostic assessment – retains a subjective element that potentially influences the final inferential outcome. Although we all know that by definition a model is a schematic and simplified approximation to reality,

"Since all models are wrong the scientist cannot obtain a *correct* one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena." [11].

We also know that "All models are wrong, but some are useful" [12], that is why model diagnostics are important. This task can be performed in both approaches. Particularly, the Bayesian framework can use predictive p–values for absolute testing [25, 1] or posterior odds ratios for relative statements [33, 38]. This is because the marginal likelihood, conditional on data, is interpreted as the evidence of the prior distribution [6].

In addition, what is objectivity in a Frequentist approach? For example, why should we use a 5% or 1% significance level rather than any other value? As someone said, the apparent objectivity is really a consensus [43]. In fact "Student" (William Gosset) saw statistical significance at any level as being "nearly valueless" in itself [65]. But, this is not just a situation in the Frequentist approach. The cut–offs given to "establish" scientific evidence against a null hypothesis in terms of log_{10} scale [34] or log_e scale [38] in Table 1.1 are also *ad hoc*.

Although the true state of nature in Bayesian inference is expressed in "degrees of belief", the distinction between the two paradigms does not reside in one being more, or less, *subjective* than the other. Rather, the differences are philosophical, pedagogical, and methodological.

2.3 Estimation, hypothesis testing and prediction

All what is required to perform estimation, hypothesis testing (model selection) and prediction in the Bayesian approach is to apply the Bayes' rule. This means coherence under a probabilistic view. But, there is no free lunch, coherence reduces flexibility. On the other hand, the Frequestist approach may be not coherent from a probabilistic point of view, but it is very flexible. This approach can be seen as a tool kit that offers inferential solutions under the umbrella of understanding probability as relative frequency. For instance, a point estimator in a Frequentist approach is found such that satisfies good sampling properties like unbiasness, efficiency, or a large sample property as consistency.

[1] Type I error is rejecting the null hypothesis when this is true, and the Type II error is not rejecting the null hypothesis when this is false.

A remarkable difference is that optimal Bayesian decisions are calculated minimizing the expected value of the loss function with respect to the posterior distribution, that is, it is conditional on observed data. On the other hand, Frequentist "optimal" actions are base on the expected values over the distribution of the estimator (a function of data) conditional on the unknown parameters, that is, it considers sampling variability.

The Bayesian approach allows to obtain the posterior distribution of any unknown object such as parameters, latent variables, future or unobserved variables or models. A nice advantage is that prediction can take into account estimation error, and predictive distributions (probabilistic forecasts) can be easily recovered.

Hypothesis testing (model selection) is based on *inductive logic* reasoning (*Inverse probability*); on the basis of what we see, we evaluate what hypothesis is most tenable, and is performed using posterior odds, which in turn are based on Bayes factors that evaluate evidence in favor of a null hypothesis taking explicitly the alternative [38], following the rules of probability [43], comparing how well the hypothesis predicts data [31], minimizing the weighted sum of type I and type II error probabilities [19, 50], and taking the implicit balance of losses [34, 9] into account. Posterior odds allows to use the same framework to analyze nested and non-nested models and perform model average. However, Bayes factors cannot be based on improper or vague priors [39], the practical interplay between model selection and posterior distributions is not as easy as it maybe in the Frequentist approach, and the computational burden can be more demanding due to solving potentially difficult integrals.

On the other hand, the Frequentist approach establishes most of its estimators as the solution of a system of equations. Observe that optimization problems reduce to solve systems. We can potentially get the distribution of these estimators, but most of the time it is needed asymptotic arguments or resampling techniques. Hypothesis testing requires pivotal quantities and/or also resampling, and prediction most of the time is based on a *plug-in approach*, which means not taking estimation error into account.[2] In addition, ancillary statistics can be used to build prediction intervals.[3] Comparing models depends on their structure, for instance, there are different Frequentist statistical approaches to compare nested and non-nested models. A nice feature in some situations is that there is a practical interplay between hypothesis testing and confidence intervals, for instance in the normal population mean hypothesis framework you cannot reject at α significance level (Type I error) any null hypothesis H_0. $\mu = \mu^0$ if μ^0 is in the $1-\alpha$ confidence interval $P(\mu \in [\hat{\mu} - |t_{N-1}^{\alpha/2}| \times \hat{\sigma}_{\hat{\mu}}, \hat{\mu} + |t_{N-1}|^{\alpha/2} \times \hat{\sigma}_{\hat{\mu}}]) = 1-\alpha$, where $\hat{\mu}$ and $\hat{\sigma}_{\hat{\mu}}$ are the maximum likelihood estimators of the mean and standard error, and $t_{N-1}^{\alpha/2}$ is

[2] A pivot quantity is a function of unobserved parameters and observations whose probability distribution does not depend on the unknown parameters.

[3] An ancillary statistic is a pivotal quantity that is also a statistic.

the quantile value of the Student's t distribution at $\alpha/2$ probability and $N-1$ degrees of freedom, N is the sample size.

A remarkable difference between the Bayesian and the Frequentist inferential frameworks is the interpretation of credible/confidence intervals. Observe that once we have estimates, such that for example the previous interval is $[0.2, 0.4]$ given a 95% confidence level, we cannot say that $P(\mu \in [0.2, 0.4]) = 0.95$ in the Frequentist framework. In fact, this probability is 0 or 1 under this approach, as μ can be there or not, the problem is that we will never know in applied settings. This due to that $P(\mu \in [\hat{\mu} - |t_{N-1}^{0.025}| \hat{\times} \sigma_{\hat{\mu}}, \hat{\mu} + |t_{N-1}^{0.025}| \times \hat{\sigma}_{\hat{\mu}}]) = 0.95$ being in the sense of repeated sampling. On the other hand, once we have the posterior distribution, we can say that $P(\mu \in [0.2, 0.4]) = 0.95$ under the Bayesian framework.

Following common practice, most of researchers and practitioners do hypothesis testing based on the p-value in the Frequentist framework. But, **what is a *p*–value?** Most of the users do not know the answer due to many times statistical inference is not performed by statisticians [7].[4] A p–value is the probability of obtaining a statistical summary of the data equal to or "more extreme" than what was actually observed, assuming that the null hypothesis is true.

Therefore, p–value calculations involve not just the observed data, but also more "extreme" hypothetical observations. So,

"What the use of p implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."[34]

It seems that common Frequentist inferential practice intertwined two different logic reasoning arguments: the p–value [22] and *significance level* [48]. The former is an informal short–run criterion, whose philosophical foundation is *reduction to absurdity*, which measures the discrepancy between the data and the null hypothesis. So, the p–value is not a direct measure of the probability that the null hypothesis is false. The latter, whose philosophical foundations is *deduction*, is based on a long–run performance such that controls the overall number of incorrect inferences in the repeated sampling without care of individual cases. The p–value fallacy consists in interpreting the p–value as the strength of evidence against the null hypothesis, and using it simultaneously with the frequency of type I error under the null hypothesis [31].

The American Statistical Association has several concerns regarding the use of the p–value as a cornerstone to perform hypothesis testing in science. This concern motivates the ASA's statement on p–values [63], which can be summarized in the following principles:

- "P–values can indicate how incompatible the data are with a specified statistical model."

[4] https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/

- "P–values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."
- "Scientific conclusions and business or policy decisions should not be based only on whether a p–value passes a specific threshold."
- "Proper inference requires full reporting and transparency."
- "A p–value, or statistical significance, does not measure the size of an effect or the importance of a result."
- "By itself, a p–value does not provide a good measure of evidence regarding a model or hypothesis."

To sum up, Fisher proposed the p-value as a witness rather than a judge. So, a p-value lower than the significance level means more inspection of the null hypothesis, but it is not a final conclusion about it.

Another difference between the Frequentists and the Bayesians is the way how scientific hypothesis are tested. The former use the p-value, whereas the latter use the Bayes factor. Observe that the p–value is associated with the probability of the data given the hypothesis, whereas the Bayes factor is associated with the probability of the hypothesis given the data. However, there is an approximate link between the t statistic and the Bayes factor for regression coefficients [52]. In particular, $|t| > (log(N)+6)^{1/2}$, corresponds to strong evidence in favor of rejecting the not relevance of a control in a regression. Observe that in this setting the threshold of the t statistic, and as a consequence the significant level, depends on the sample size. Observe that this setting agrees with the idea in experimental designs of selecting the sample size such that we control Type I and Type II errors. In observational studies we cannot control the sample size, but we can select the significance level.

See also [56] and [5] for nice exercises to reveal potential flaws of the p–value (p) due to $p \sim U[0,1]$ under the null hypothesis,[5] and calibrations of the p-value to interpret them as the odds ratio and the error probability. In particular, $B(p) = -e \times p \times \log(p)$ when $p < e^{-1}$, and interpret this as the Bayes factor of H_0 to H_1, where H_1 denotes the unspecified alternative to H_0, and $\alpha(p) = (1 + [-e \times p \times \log(p)]^{-1})^{-1}$ as the error probability α in rejecting H_0. Take into account that $B(p)$ and $\alpha(p)$ are lower bounds.

Logic of argumentation in the Frequentist approach is based on *deductive logic*, this means that it starts from a statement about the true state of nature (null hypothesis), and predicts what should be seen if this statement were true. On the other hand, the Bayesian approach is based on *inductive logic*, this means that it defines what hypothesis is more consistent with what is seen. The former inferential approach establishes that the true of the premises implies the true of the conclusion, that is why we reject or not reject hypothesis. The latter establishes that the premises supply some evidence, but not

[5] https://joyeuserrance.wordpress.com/2011/04/22/proof-that-p-values-under-the-null-are-uniformly-distributed/ for a simple proof.

full assurance, of the true of the conclusion, that is why we get probabilistic statements.

Here, there is a difference between *effects of causes* (forward causal inference) and *causes of effects* (reverse causal inference) [28, 17]. To illustrate this point, imagine that a firm increases the price of a specific good, then economic theory would say that its demand decreases. The premise (null hypothesis) is a price increase, and the consequence is a demand reduction. Another view would be to observe a demand reduction, and try to identify which cause is more tenable. For instance, demand reduction can be caused by any positive supply shocks or any negative demand shocks. The Frequentist logic sees the first view, and the Bayesian reasoning gives the probability associated with possible causes.

2.4 The likelihood principle

The **likelihood principle** states that in making inference or decisions about the state of the nature all the relevant *experimental* information is given by the *likelihood function*. The Bayesian framework follows this statement, that is, it is conditional on observed data.

We follow [6], who in turns followed [44], to illustrate the likelihood principle. We are given a coin such that we are interested in the probability, θ, of having it come up heads when flipped. It is desired to test H_0. $\theta = 1/2$ versus H_1. $\theta > 1/2$. An experiment is conducted by flipping the coin (independently) in a series of trials, the results of which is the observation of 9 heads and 3 tails.

This is not yet enough information to specify $p(y|\theta)$, since the series of trials was not explained. Two possibilities:

1. The experiment consisted of a predetermine 12 flips, so that $Y = [Heads]$ would be $\mathcal{B}(12, \theta)$, then $p_1(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = 220 \times \theta^9(1-\theta)^3$.
2. The experiment consisted of flipping the coin until 3 tails were observed ($r = 3$). Then, Y, the number of failures (heads) until getting 3 tails, is $\mathcal{NB}(3, 1-\theta)$. Then, $p_2(y|\theta) = \binom{y+r-1}{r-1}(1-(1-\theta)^y(1-\theta)^r = 55 \times \theta^9(1-\theta)^3$.

Using a Frequentist approach, the significance level of $y = 9$ using the Binomial model against $\theta = 1/2$ would be:

$$\alpha_1 = P_{1/2}(Y \geq 9) = p_1(9|1/2) + p_1(10|1/2) + p_1(11|1/2) + p_1(12|1/2) = 0.073.$$

R code. The likelihood principle: Binomial model

```
success <- 9
# Number of observed success in n trials
n <- 12
# Number of trials
siglevel <- sum(sapply(9:n,function(y)dbinom(y,n,0.5)))
siglevel
0.073
```

For the Negative Binomial model, the significance level would be:

$$\alpha_2 = P_{1/2}(Y \geq 9) = p_2(9|1/2) + p_2(10|1/2) + \ldots = 0.0327.$$

R code. The likelihood principle: Negative Binomial model

```
success <- 3
# Number of target success (tails)
failures <- 9
# Number of failures
siglevel <- 1 - pnbinom((failures - 1),success,0.5)
siglevel
0.0327
```

We arrive to different conclusions using a significance level equal to 5%, whereas we obtain the same outcomes using a Bayesian approach because the kernels of both distributions are the same ($\theta^9 \times (1-\theta)^3$).

2.5 Why is not the Bayesian approach that popular?

At this stage, we may wonder why the Bayesian statistical framework is not the dominant inferential approach despite that it has its historical origin in 1763 [4], whereas the Frequentist statistical framework was largely developed in the early 20th century. The scientific battle over the Bayesian inferential approach lasted for 150 years, and this maybe explained by some of the following facts.

There is an issue regarding *apparent subjectivity* as the Bayesian inferential approach runs counter the strong conviction that science demands objectivity, and Bayesian probability is a measure of degrees of belief, where the initial prior maybe just a guess; this was not accepted as objective and rigorous science. Initial critics said that Bayes was quantifying ignorance as he set equal probabilities to any potential result. As a consequence, prior distributions were damned [46].

Bayes himself seemed not to have believed in his idea. Although, it seems that Bayes achieved his breakthrough during the late 1740s, he did not send it off to the Royal Society for publication. It was his friend, Richard Price, another Presbyterian minister, who rediscovered Bayes' idea, polished it and published.

However, it was Laplace who independently generalized Bayes' theorem in 1781. He used it initially in gambling problems, and soon after in astronomy, mixing different sources of information in order to leverage research in specific situations where data was scarce. Then, he wanted to use his discovery to find the probability of causes, and thought that this required large data sets, and turned into demography. In this field, he had to perform large calculations that demanded to develop smart approximations, creating the Laplace's approximation and the central limit theorem [40]; although, at the cost of apparently leaving his research on Bayesian inference.

Once *Laplace was gone in 1827*, the Bayes' rule disappeared from the scientific spectrum for almost a century. In part, personal attacks against Laplace made the rule be forgotten, and also, the old fashion thought that statistics does not have to say anything about causation, and that the prior is very subjective to be compatible with science. Although, practitioners used it to solve problems in astronomy, communication, medicine, military and social issues with remarkable results.

Thus, the concept of degrees of belief to operationalize probability was gone in name of scientific objectivity, and probability as the frequency an event occurs in many repeatable trials became the rule. Laplace critics argued that those concepts were diametric opposites, although, Laplace considered them as basically equivalent when large sample sizes are involved [46].

The era of the Frequentists or sampling theorists began, lead by Karl Pearson, and his nemesis, Ronald Fisher, both brilliant, against the inverse probability approach, persuasive and dominant characters that made impossible to argue against their ideas. Karl Pearson legacy was taken by his son Egon, and Egon's friend Neyman, both inherited the anti-Bayesian and anti-Fisher legacy.

Despite the anti-Bayesian campaign among statisticians, there were some independent characters developing Bayesian ideas, Borel, Ramsey and de Fineti, all of them isolated in different countries, France, England and Italy. However, the anti-Bayesian trio of Fisher, Neyman and Egon Person got all the attention during the 1920s and 1930s. Only, a geophysicist, Harold Jeffreys, kept alive Bayesian inference in the 1930s and 1940s. Jeffreys was a

very quiet, shy, uncommunicative gentleman working at Cambridge in the astronomy department. He was Fisher's friend thanks to his character, although they were diametric opposites regarding the Bayesian inferential approach, facing very high intellectual battles. Unfortunately for the Bayesian approach, *Jeffreys lost*, he was very technical using confusing high level mathematics, worried about inference from scientific evidence, not guiding future actions based on decision theory, which was very important in that era for mathematical statistics due to the Second World War. On the other hand, Fisher was a very dominant character, persuasive in public and a master of practice, his techniques were written in a popular style with minimum mathematics.

However, Bayes' rule achieved remarkable results in applied settings like the AT&T company or the social security system in USA. Bayesian inference also had a relevant role during the second World War and the Cold War. Alan Turing used inverse probability at Bletchley Park to crack German messages called Enigma code used by U-boats, Andrei Kolmogorov used it to improved firing tables of Russia's artillery, Bernard Koopman applied it for searching targets in the open sea and the RAND Corporation used it in the Cold War. Unfortunately, *these Bayesian developments were top secrets for almost 40 years* that keep classified the contribution of inverse probability in modern human history.

During 1950s and 1960s three mathematicians lead the rebirth of the Bayesian approach, Good, Savage and Lindley. However, it seems that they were unwilling to apply their theories to real problems, and despite that the Bayesian approach proved its worth, for instance, in business decisions, navy search, lung cancer, etc, it was applied to simple models due to its *mathematical complexity and requirement of large computations*. But, there were some breakthrough that change this. First, hierarchical models introduced by Lindley and Smith, where a complex model is decomposed into many easy to solve models, and second, Markov chain Monte Carlo methods developed by Hastings in the 1970s [32] and the Geman brothers in the 1980s [29]. These methods were introduced into the Bayesian inferential framework in the 1990s by Gelfand and Smith [23], and Tierney [61], when desktop computers got enough computational power to solve complex models. Since then, the Bayesian inferential framework has gained increasing popularity among practitioners and scientists.

2.6 A simple working example

We will illustrate some conceptual differences between the Bayesian and Frequentist statistical approaches performing inference given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]$, where $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, N$.

In particular, we set $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma) \propto \frac{1}{\sigma}$. This is a standard *non-*

informative improper prior (Jeffreys prior, see Chapter **??**), that is, this prior is perfectly compatible with sample information. In addition, we are assuming independent priors for μ and σ. Then,

$$\begin{aligned}
\pi(\mu,\sigma|\mathbf{y}) &\propto \frac{1}{\sigma}\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}\\
&= \frac{1}{\sigma}\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}((y_i-\bar{y})-(\mu-\bar{y}))^2\right\}\\
&= \frac{1}{\sigma}\exp\left\{-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right\}\times(\sigma)^{-N}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2\right\}\\
&= \frac{1}{\sigma}\exp\left\{-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right\}\times(\sigma)^{-(\alpha_n+1)}\exp\left\{-\frac{\alpha_n\hat{\sigma}^2}{2\sigma^2}\right\},
\end{aligned}$$

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

The first term in the last expression is the kernel of a normal density, $\mu|\sigma,\mathbf{y}\sim N(\bar{y},\sigma^2/N)$. The second term is the kernel of an inverted gamma density [64], $\sigma|\mathbf{y}\sim IG(\alpha_n,\hat{\sigma}^2)$. Therefore, $\pi(\mu|\sigma,\mathbf{y})=(2\pi\sigma^2/N)^{-1/2}\exp\left\{\frac{-N}{2\sigma^2}(\mu-\bar{y})^2\right\}$ and $\pi(\sigma|\mathbf{y})=\frac{2}{\Gamma(\alpha_n/2)}\left(\frac{\alpha_n\hat{\sigma}^2}{2}\right)^{\alpha_n/2}\frac{1}{\sigma^{\alpha_n+1}}\times\exp\left\{-\frac{\alpha_n\hat{\sigma}^2}{2\sigma^2}\right\}$.

Observe that $\mathbb{E}[\mu|\sigma,\mathbf{y}]=\bar{y}$, this is also the maximum likelihood (Frequentist) point estimate of μ in this setting. In addition, the Frequentist $(1-\alpha)\%$ confidence interval and the Bayesian $(1-\alpha)\%$ credible interval have exactly the same form, $\bar{y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$, where $z_{\alpha/2}$ is the $\alpha/2$ percentile of a standard normal distribution. However, the interpretations are totally different. The confidence interval has a probabilistic interpretation under sampling variability of $\bar{Y}$, that is, in repeated sampling $(1-\alpha)\%$ of the intervals $\bar{Y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$ would include μ, but given an observed realization of $\bar{Y}$, say $\bar{y}$, the probability of $\bar{y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$ including μ is 1 or 0, that is why we say a $(1-\alpha)\%$ confidence interval. On the other hand, $\bar{y}\pm|z_{\alpha/2}|\frac{\sigma}{\sqrt{N}}$ has a simple probabilistic interpretation in the Bayesian framework, there is a $(1-\alpha)\%$ probability that μ lies in this interval.

If we want to get the marginal posterior density of μ,

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

The fourth line is due to having the kernel of a inverted gamma density with N degrees of freedom in the integral [64].

The last expression is the kernel of a Student's t density function with $\alpha_n = N-1$ degrees of freedom, expected value equal to $\bar{y}$, and variance $\frac{\hat{\sigma}^2}{N}\left(\frac{\alpha_n}{\alpha_n-2}\right)$. Then, $\mu|\mathbf{y} \sim t\left(\bar{y}, \frac{\hat{\sigma}^2}{N}\left(\frac{\alpha_n}{\alpha_n-2}\right), \alpha_n\right)$.

Observe that a $(1-\alpha)\%$ confidence interval and $(1-\alpha)\%$ credible interval have exactly the same expression, $\bar{y} \pm |t^{\alpha_n}_{\alpha/2}|\frac{\hat{\sigma}}{\sqrt{N}}$, where $t^{\alpha_n}_{\alpha/2}$ is the $\alpha/2$ percentile of a Student's t distribution. But again, the interpretations are totally different.

The mathematical similarity between the Frequentist and Bayesian expressions in this example is due to using an improper prior.

2.6.1 Example: Math test

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. Assuming an improper prior equal to $1/\sigma$,

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

R code. Example: Math test

```
N <- 50 # Sample size
y_bar <- 102 # Sample mean
s2 <- 10 # Sample variance
alpha <- N - 1
serror <- (s2/N)^0.5
LimInf <- y_bar - abs(qt(0.025, alpha)) * serror
LimInf
101.101
# Lower bound
LimSup <- y_bar + abs(qt(0.025, alpha)) * serror
LimSup
102.898
# Upper bound
y.cut <- 103
P <- 1-metRology::pt.scaled(y.cut, df = alpha,
mean = y_bar, sd = serror)
P
0.0149
# Probability of mu greater than y.cut
```

2.7 Summary: Chapter 2

The differences between the Bayesian and Frequentist inferential approaches are philosophical, including as pertains to the role of probability; pedagogical, in particular as relates to the use of inference to inform decision making; and methodological, as having differences in their mathematical and computational frameworks. Although at methodological level, the debate has become considerably muted, except for some aspects of inference, with the recognition that each approach has a great deal to contribute to statistical practice [30, 2, 37]. As Bradley Efron said "Computer-age statistical inference at its most successful **combines** elements of the two philosophies" [21].

2.8 Exercises: Chapter 2

1. **Jeffreys-Lindley's paradox**

The **Jeffreys-Lindley's paradox** [34, 45] 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 0.0235 using the normal approximation, $Y \sim \mathcal{N}(N\times\theta, N\times\theta\times(1-\theta))$.

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

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

   Show that

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

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

$$
\begin{aligned}
PO_{01} &= \frac{p(\mathbf{y}|\mathcal{M}_0)}{p(\mathbf{y}|\mathcal{M}_1)} \\
&= \frac{\Gamma((\alpha_n+1)/2)}{\Gamma(1/2)\Gamma(\alpha_n/2)}(\alpha_n\hat{\sigma}^2/N)^{-1/2}\left[1+\frac{(\mu_0-\bar{y})^2}{\alpha_n\hat{\sigma}^2/N}\right]^{-\left(\frac{\alpha_n+1}{2}\right)},
\end{aligned}
$$

   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.

3. Using the setting of the **Example: Math test** in subsection 2.6.1, 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.

3

Objective and subjective Bayesian approaches

4

Cornerstone models: Conjugate families

We will introduce conjugate families in basic statistical models with examples, solving them analytically and computationally using R. We will have some mathematical, and computational exercises in R.

4.1 Motivation of conjugate families

Observing the three fundamental pieces of Bayesian analysis: the posterior distribution (parameter inference), the marginal likelihood (hypothesis testing), and the predictive distribution (prediction), equations 4.1, 4.2 and 4.3, respectively,

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) \times \pi(\theta)}{p(\mathbf{y})}, \tag{4.1}$$

$$p(\mathbf{y}) = \int_{\mathbf{\Theta}} p(\mathbf{y}|\theta)\pi(\theta)d\theta, \tag{4.2}$$

and

$$p(\mathbf{Y}_0|\mathbf{y}) = \int_{\mathbf{\Theta}} p(\mathbf{Y}_0|\theta)\pi(\theta|\mathbf{y})d\theta, \tag{4.3}$$

we can understand that some of the initial limitations of the application of the Bayesian analysis were associated with the absence of algorithms to draw from non-standard posterior distributions (equation 4.1), and the lack of analytical solutions of the marginal likelihood (equation 4.2) and the predictive distribution (equation 4.3). Both issues requiring computational power.

Although there were algorithms to sample from non-standard posterior distributions since the second half of the last century [47, 32, 29], their particular application in the Bayesian framework emerged later [23, 61], maybe until the increasing computational power of desktop computers. However, it is also common practice nowadays to use models that have standard conditional posterior distributions to mitigate computational requirements. In addition, nice mathematical tricks plus computational algorithms [24, 14, 15] and approximations [62, 35] are used to obtain the marginal likelihood (prior predictive).

Despite these advances, there are two potentially conflicting desirable model specification features that we can see from equations 4.1, 4.2 and 4.3: analytical solutions and the posterior distribution in the same family as the prior distribution for a given likelihood. The latter is called *conjugate priors*, a family of priors that is closed under sampling [55].

These features are desirable as the former implies facility to perform hypothesis testing and predictive analysis, and the latter means invariance of the prior-to-posterior updating. Both features imply less computational burden.

We can easily achieve each of these features independently, for instance using improper priors for analytical tractability, and defining in a broad sense the family of prior distributions for prior conjugacy. However, these features are in conflict.

Fortunately, we can achieve these two nice characteristics if we assume that the data generating process is given by a distribution function in the *exponential family*. That is, given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$, a probability density function $p(\mathbf{y}|\theta)$ belongs to the exponential family if it has the form

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{N} h(y_i) C(\theta) \exp\left\{\eta(\theta)^\top \mathbf{T}(y_i)\right\} \\ &= h(\mathbf{y}) C(\theta)^N \exp\left\{\eta(\theta)^\top \mathbf{T}(\mathbf{y})\right\} \\ &= h(\mathbf{y}) \exp\left\{\eta(\theta)^\top \mathbf{T}(\mathbf{y}) - A(\theta)\right\}, \end{aligned} \tag{4.4}$$

where $h(\mathbf{y}) = \prod_{i=1}^{N} h(y_i)$ is a non-negative function, $\eta(\theta)$ is a known function of the parameters, $A(\theta) = \log\left\{\int_{\mathbf{Y}} h(\mathbf{y}) \exp\left\{\eta(\theta)^\top \mathbf{T}(\mathbf{y})\right\} d\mathbf{y}\right\} = -N \log(C(\theta))$ is a normalization factor, and $\mathbf{T}(\mathbf{y}) = \sum_{i=1}^{N} \mathbf{T}(y_i)$ is the vector of sufficient statistics of the distribution (by the factorization theorem).

If the support of $\mathbf{y}$ is independent of θ, then the family is said to be *regular*, otherwise it is irregular. In addition, if we set $\eta = \eta(\theta)$, then the exponential family is said to be in the *canonical form*

$$\begin{aligned} p(\mathbf{y}|\theta) &= h(\mathbf{y}) D(\eta)^N \exp\left\{\eta^\top \mathbf{T}(\mathbf{y})\right\} \\ &= h(\mathbf{y}) \exp\left\{\eta^\top \mathbf{T}(\mathbf{y}) - B(\eta)\right\}. \end{aligned}$$

A nice feature of this representation is that $\mathbb{E}[\mathbf{T}(\mathbf{y})|\eta] = \nabla B(\eta)$ and $Var[\mathbf{T}(\mathbf{y})|\eta] = \nabla^2 B(\eta)$.

4.1.1 Examples of exponential family distributions

1. **Discrete distributions**

   Poisson distribution

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a *Poisson distribution* let's show that $p(\mathbf{y}|\lambda)$ is in the exponential family.

$$\begin{aligned} p(\mathbf{y}|\lambda) &= \prod_{i=1}^{N} \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \\ &= \frac{\lambda^{\sum_{i=1}^N y_i} \exp(-N\lambda)}{\prod_{i=1}^N y_i!} \\ &= \frac{\exp(-N\lambda) \exp(\sum_{i=1}^N y_i \log(\lambda))}{\prod_{i=1}^N y_i!}, \end{aligned}$$

then $h(\mathbf{y}) = \left[\prod_{i=1}^N y_i!\right]^{-1}$, $\eta(\lambda) = \log(\lambda)$, $T(\mathbf{y}) = \sum_{i=1}^N y_i$ (sufficient statistic) and $C(\lambda) = \exp(-\lambda)$.
If we set $\eta = \log(\lambda)$, then

$$p(\mathbf{y}|\eta) = \frac{\exp(\eta \sum_{i=1}^N y_i - N \exp(\eta))}{\prod_{i=1}^N y_i!},$$

such that $B(\eta) = N\exp(\eta)$, then $\nabla(B(\eta)) = N\exp(\eta) = N\lambda = \mathbb{E}\left[\sum_{i=1}^N y_i \middle| \lambda\right]$, that is, $\mathbb{E}\left[\frac{\sum_{i=1}^N y_i}{N} \middle| \lambda\right] = \mathbb{E}[\bar{y}|\lambda] = \lambda$, and $\nabla^2(B(\eta)) = N\exp(\eta) = N\lambda = Var\left[\sum_{i=1}^N y_i \middle| \lambda\right] = N^2 \times Var\left[\bar{y}|\lambda\right]$, then $Var\left[\bar{y}|\lambda\right] = \frac{\lambda}{N}$.

Bernoulli distribution

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a **Bernoulli distribution** let's show that $p(\mathbf{y}|\theta)$ is in the exponential family.

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

then $h(\mathbf{y}) = \mathbb{I}[y_i \in \{0,1\}]$, $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, $T(\mathbf{y}) = \sum_{i=1}^N y_i$ and $C(\theta) = 1-\theta$.

Write this distribution in the canonical form, and find the mean and variance of the sufficient statistic (Exercise 1).

Multinomial distribution

Given a random sample $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$ from a *m-dimensional multinomial distribution*, where $\mathbf{y}_i = [y_{i1}, \ldots, y_{im}]$, $\sum_{l=1}^{m} y_{il} = n$, n independent trials each of which leads to a success for exactly one of m categories with probabilities $\theta = [\theta_1, \theta_2, \ldots, \theta_m]$, $\sum_{l=1}^{m} \theta_l = 1$. Let's show that $p(\mathbf{y}|\theta)$ is in the exponential family.

$$
\begin{aligned}
p(\mathbf{y}|\theta) &= \prod_{i=1}^{N} \frac{n!}{\prod_{l=1}^{m} y_{il}!} \prod_{l=1}^{m} \theta_l^{y_{il}} \\
&= \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!} \exp\left\{ \sum_{i=1}^{N}\sum_{l=1}^{m} y_{il} \log(\theta_l) \right\} \\
&= \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!} \exp\left\{ \left(N \times n - \sum_{i=1}^{N}\sum_{l=1}^{m-1} y_{il} \right) \log(\theta_m) \right. \\
&\left. + \sum_{i=1}^{N}\sum_{l=1}^{m-1} y_{il} \log(\theta_l) \right\} \\
&= \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!} \theta_m^{N\times n} \exp\left\{ \sum_{i=1}^{N}\sum_{l=1}^{m-1} y_{il} \log(\theta_l/\theta_m) \right\},
\end{aligned}
$$

then $h(\mathbf{y}) = \frac{(n!)^N}{\prod_{i=1}^{N}\prod_{l=1}^{m} y_{il}!}$, $\eta(\theta) = \left[\log\left(\frac{\theta_1}{\theta_m}\right) \ldots \log\left(\frac{\theta_{m-1}}{\theta_m}\right)\right]$, $T(\mathbf{y}) = \left[\sum_{i=1}^{N} y_{i1} \cdots \sum_{i=1}^{N} y_{im-1}\right]$ and $C(\theta) = \theta_m^n$.

2. **Continuous distributions**

Normal distribution

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a *normal distribution* let's show that $p(\mathbf{y}|\mu, \sigma^2)$ is in the exponential family.

$$
\begin{aligned}
p(\mathbf{y}|\mu,\sigma^2) &= \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i-\mu\right)^2\right\} \\
&= (2\pi)^{-N/2}(\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-\mu\right)^2\right\} \\
&= (2\pi)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N} y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{N} y_i \right. \\
&\left. - N\frac{\mu^2}{2\sigma^2} - \frac{N}{2}\log(\sigma^2)\right\},
\end{aligned}
$$

then $h(\mathbf{y}) = (2\pi)^{-N/2}$, $\eta(\mu,\sigma^2) = \left[\frac{\mu}{\sigma^2} \ \frac{-1}{2\sigma^2}\right]$, $T(\mathbf{y}) = \left[\sum_{i=1}^{N} y_i \ \sum_{i=1}^{N} y_i^2\right]$ and $C(\mu,\sigma^2) = \exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}\right\}$.
Observe that

$$
p(\mathbf{y}|\mu,\sigma^2) = (2\pi)^{-N/2} \exp\left\{\eta_1 \sum_{i=1}^{N} y_i + \eta_2 \sum_{i=1}^{N} y_i^2 - \frac{N}{2}\log(-2\eta_2) + \frac{N}{4}\frac{\eta_1^2}{\eta_2}\right\},
$$

where $B(\eta) = \frac{N}{2}\log(-2\eta_2) - \frac{N}{4}\frac{\eta_1^2}{\eta_2}$. Then,

$$
\nabla B(\eta) = \begin{bmatrix} -\frac{N}{2}\frac{\eta_1}{\eta_2} \\ -\frac{N}{2}\frac{1}{\eta_2} + \frac{N}{4}\frac{\eta_1^2}{\eta_2^2} \end{bmatrix} = \begin{bmatrix} N\times\mu \\ N\times(\mu^2+\sigma^2) \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[\sum_{i=1}^{N} y_i \middle| \mu,\sigma^2\right] \\ \mathbb{E}\left[\sum_{i=1}^{N} y_i^2 \middle| \mu,\sigma^2\right] \end{bmatrix}.
$$

Multivariate normal distribution

Given $\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \ldots \ \mathbf{y}_p]$ a $N \times p$ matrix such that $\mathbf{y}_i \sim N_p(\mu, \mathbf{\Sigma})$, $i = 1, 2, \ldots, N$, that is, each i-th row of $\mathbf{Y}$ follows a *multivariate normal distribution*. Then, assuming independence between rows, let's show that $p(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N|\mu, \mathbf{\Sigma})$ is in the exponential family.

$$
\begin{aligned}
p(\mathbf{y}_1,\ldots,\mathbf{y}_N|\mu,\mathbf{\Sigma}) &= \prod_{i=1}^{N}(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_i-\mu\right)^\top\mathbf{\Sigma}^{-1}\left(\mathbf{y}_i-\mu\right)\right\}\\
&= (2\pi)^{-pN/2}|\Sigma|^{-N/2}\exp\left\{-\frac{1}{2}tr\left[\sum_{i=1}^{N}\left(\mathbf{y}_i-\mu\right)^\top\mathbf{\Sigma}^{-1}\left(\mathbf{y}_i-\mu\right)\right]\right\}\\
&= (2\pi)^{-pN/2}|\Sigma|^{-N/2}\exp\left\{-\frac{1}{2}tr\left[\left(\mathbf{S}+N\left(\mu-\hat{\mu}\right)\left(\mu-\hat{\mu}\right)^\top\right)\mathbf{\Sigma}^{-1}\right]\right\}\\
&= (2\pi)^{-pN/2}\exp\left\{-\frac{1}{2}\left[\left(vec\left(\mathbf{S}\right)^\top+Nvec\left(\hat{\mu}\hat{\mu}^\top\right)^\top\right)vec\left(\mathbf{\Sigma}^{-1}\right)\right.\right.\\
&\left.\left.-2N\hat{\mu}^\top vec\left(\mu^\top\mathbf{\Sigma}^{-1}\right)+Ntr\left(\mu\mu^\top\mathbf{\Sigma}^{-1}\right)+N\log(|\mathbf{\Sigma}|)\right]\right\},
\end{aligned}
$$

where the second line uses the trace operator (tr), and its invariability under cyclic permutation is used in the third line. In addition, we add and subtract $\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i$ in each parenthesis such that we get $\mathbf{S}=\sum_{i=1}^{N}\left(\mathbf{y}_i-\hat{\mu}\right)\left(\mathbf{y}_i-\hat{\mu}\right)^\top$. We get the fourth line after using some properties of the trace operator to introduce the vectorization operator (vec), and collecting terms.

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

4.2 Conjugate prior to exponential family

Theorem 4.2.1

The prior distribution $\pi(\theta)\propto C(\theta)^{b_0}\exp\left\{\eta(\theta)^\top\mathbf{a}_0\right\}$ is conjugate to the exponential family (equation 4.4).

Proof

$$
\begin{aligned}
\pi(\theta|\mathbf{y}) &\propto C(\theta)^{b_0}\exp\left\{\eta(\theta)^\top\mathbf{a}_0\right\}\times h(\mathbf{y})C(\theta)^N\exp\left\{\eta(\theta)^\top\mathbf{T}(\mathbf{y})\right\}\\
&\propto C(\theta)^{N+b_0}\exp\left\{\eta(\theta)^\top\left(\mathbf{T}(\mathbf{y})+\mathbf{a}_0\right)\right\}.
\end{aligned}
$$

Observe that the posterior is in the exponential family, $\pi(\theta|\mathbf{y})\propto C(\theta)^{\beta_n}\exp\left\{\eta(\theta)^\top\alpha_n\right\}$, $\beta_n=N+b_0$ and $\alpha_n=\mathbf{T}(\mathbf{y})+\mathbf{a}_0$.

Remarks

We see comparing the prior and the likelihood that b_0 plays the role of a hypothetical sample size, and $\mathbf{a}_0$ plays the role of hypothetical sufficient statistics. This view helps the elicitation process.

In addition, we established the result in the *standard form* of the exponential family. We can also establish this result in the *canonical form* of the exponential family. Observe that given $\eta = \eta(\theta)$, another way to get a prior for η is to use the change of variable theorem given a bijective function.

In the setting where there is a regular conjugate prior, [20] show that we obtain a posterior expectation of the sufficient statistics that is a weighted average between the prior expectation and the likelihood estimate.

4.2.1 Examples: Theorem 4.2.1

1. **Likelihood functions from discrete distributions**

   The Poisson-gamma model

   Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a Poisson distribution then a conjugate prior density for λ has the form

$$\begin{aligned}\pi(\lambda) &\propto (\exp(-\lambda))^{b_0} \exp\{a_0 \log(\lambda)\} \\ &= \exp(-\lambda b_0)\lambda^{a_0} \\ &= \exp(-\lambda\beta_0)\lambda^{\alpha_0 - 1}.\end{aligned}$$

   This is the kernel of a gamma density in the *rate parametrization*, $G(\alpha_0, \beta_0)$, $\alpha_0 = a_0 + 1$ and $\beta_0 = b_0$.[1] Then, a prior conjugate distribution for the Poisson likelihood is a gamma distribution.

   Taking into account that $\sum_{i=1}^N y_i$ is a sufficient statistic for the Poisson distribution, then we can think about a_0 as the number of occurrences in b_0 experiments. Observe that

$$\begin{aligned}\pi(\lambda|\mathbf{y}) &\propto \exp(-\lambda\beta_0)\lambda^{\alpha_0 - 1} \times \exp(-N\lambda)\lambda^{\sum_{i=1}^N y_i} \\ &= \exp(-\lambda(N + \beta_0))\lambda^{\sum_{i=1}^N y_i + \alpha_0 - 1}.\end{aligned}$$

   As expected, this is the kernel of a gamma distribution, which means $\lambda|\mathbf{y} \sim G(\alpha_n, \beta_n)$, $\alpha_n = \sum_{i=1}^N y_i + \alpha_0$ and $\beta_n = N + \beta_0$.

   Observe that α_0/β_0 is the prior mean, and α_0/β_0^2 is the prior variance. Then, $\alpha_0 \to 0$ and $\beta_0 \to 0$ imply a non-informative prior such that the posterior mean converges to the maximum likelihood estimator $\bar{y} = \frac{\sum_{i=1}^N y_i}{N}$,

[1] Another parametrization of the gamma density is the *scale parametrization* where $\kappa_0 = 1/\beta_0$. See the health insurance example in Chapter 1.

$$\begin{aligned}\mathbb{E}[\lambda|\mathbf{y}] &= \frac{\alpha_n}{\beta_n} \\ &= \frac{\sum_{i=1}^{N} y_i + \alpha_0}{N + \beta_0} \\ &= \frac{N\bar{y}}{N + \beta_0} + \frac{\alpha_0}{N + \beta_0}.\end{aligned}$$

The posterior mean is a weighted average between sample and prior information. This is a general result from regular conjugate priors [20]. Observe that $\mathbb{E}[\lambda|\mathbf{y}] = \bar{y}, \lim N \to \infty$.

In addition, $\alpha_0 \to 0$ and $\beta_0 \to 0$ corresponds to $\pi(\lambda) \propto \frac{1}{\lambda}$, which is an improper prior. Improper priors have bad consequences on Bayes factors (hypothesis testing), see bellow a discussion of this in the linear regression framework. In this setting, we can get analytical solutions for the marginal likelihood and the predictive distribution (see the health insurance example and Exercise 3 in Chapter 1).

The Bernoulli-beta model

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a Bernoulli distribution then a conjugate prior density for θ has the form

$$\begin{aligned}\pi(\theta) &\propto (1-\theta)^{b_0} \exp\left\{a_0 \log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &= (1-\theta)^{b_0 - a_0}\theta^{a_0} \\ &= \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}.\end{aligned}$$

This is the kernel of a beta density, $B(\alpha_0, \beta_0)$, $\alpha_0 = a_0 + 1$ and $\beta_0 = b_0 - a_0 + 1$. A prior conjugate distribution for the Bernoulli likelihood is a beta distribution. Given that b_0 is the hypothetical sample size, and a_0 is the hypothetical sufficient statistic, which is the number of successes, then $b_0 - a_0$ is the number of failures. This implies that α_0 is the number of prior successes plus one, and β_0 is the number of prior failures plus one. Given that the mode of a beta distributed random variable is $\frac{\alpha_0 - 1}{\alpha_0 + \beta_0 - 2} = \frac{a_0}{b_0}$, then we have the prior probability of success. Setting $\alpha_0 = 1$ and $\beta_0 = 1$, which implies a 0-1 uniform distribution, corresponds to a setting with 0 successes (and 0 failures) in 0 experiments.

Observe that

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

The posterior distribution is beta, $\theta|\mathbf{y} \sim B(\alpha_n, \beta_n)$, $\alpha_n = \alpha_0 + \sum_{i=1}^N y_i$ and $\beta_n = \beta_0 + N - \sum_{i=1}^N y_i$, where the posterior mean $\mathbf{E}[\theta|\mathbf{y}] = \frac{\alpha_n}{\alpha_n+\beta_n} = \frac{\alpha_0+N\bar{y}}{\alpha_0+\beta_0+N} = \frac{\alpha_0+\beta_0}{\alpha_0+\beta_0+N}\frac{\alpha_0}{\alpha_0+\beta_0} + \frac{N}{\alpha_0+\beta_0+N}\bar{y}$. The posterior mean is a weighted average between the prior mean and the maximum likelihood estimate.

El marginal likelihood in this setting is

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

where $B(\cdot,\cdot)$ is the beta function.

In addition, the predictive density is

$$\begin{aligned}p(Y_0|\mathbf{y}) &= \int_0^1 \theta^{y_0}(1-\theta)^{1-y_0} \times \frac{\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}}{B(\alpha_n,\beta_n)} d\theta \\ &= \frac{B(\alpha_n+y_0, \beta_n+1-y_0)}{B(\alpha_n,\beta_n)} \\ &= \frac{\Gamma(\alpha_n+\beta_n)\Gamma(\alpha_n+y_0)\Gamma(\beta_n+1-y_0)}{\Gamma(\alpha_n+\beta_n+1)\Gamma(\alpha_n)\Gamma(\beta_n)} \\ &= \left\{\begin{array}{ll} \frac{\alpha_n}{\alpha_n+\beta_n}, & y_0 = 1 \\ \frac{\beta_n}{\alpha_n+\beta_n}, & y_0 = 0 \end{array}\right\}.\end{aligned}$$

This is a Bernoulli distribution with probability of success equal to $\frac{\alpha_n}{\alpha_n+\beta_n}$.

The multinomial-Dirichlet model

Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a multinomial distribution then a conjugate prior density for $\theta = [\theta_1, \theta_2, \ldots, \theta_m]$ has the form

$$\begin{aligned}\pi(\theta) &\propto \theta_m^{b_0} \exp\left\{\eta(\theta)^\top \mathbf{a}_0\right\} \\ &= \prod_{l=1}^{m-1} \theta_l^{a_{0l}} \theta_m^{b_0 - \sum_{l=1}^{m-1} a_{0l}} \\ &= \prod_{l=1}^{m} \theta_l^{\alpha_{0l}-1},\end{aligned}$$

where $\eta(\theta) = \left[\log\left(\frac{\theta_1}{\theta_m}\right), \ldots, \log\left(\frac{\theta_{m-1}}{\theta_m}\right)\right]$, $\mathbf{a}_0 = [a_{01}, \ldots, a_{am-1}]^\top$, $\alpha_0 = [\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0m}]$, $\alpha_{0l} = a_{0l} + 1$, $l = 1, 2, \ldots, m-1$ and $\alpha_{0m} = b_0 - \sum_{l=1}^{m-1} a_{0l} + 1$.

This is the kernel of a Dirichlet distribution, that is, the prior distribution is $D(\alpha_0)$.

Observe that a_{0l} is the number of hypothetical number of times outcome l is observed over the hypothetical b_0 trials. Setting $\alpha_{0l} = 1$, that is a uniform distribution over the open standard simplex, implicitly we set $a_{0l} = 0$, which means that there are 0 occurrences of category l in $b_0 = 0$ experiments.

The posterior distribution of the multinomial-Dirichlet model is given by

$$\begin{aligned}\pi(\theta|\mathbf{y}) &\propto \prod_{l=1}^{m} \theta_l^{\alpha_{0l}-1} \times \prod_{l=1}^{m} \theta_l^{\sum_{i=1}^{N} y_{il}} \\ &= \prod_{l=1}^{m} \theta_l^{\alpha_{0l} + \sum_{i=1}^{N} y_{il} - 1}.\end{aligned}$$

This is the kernel of a Dirichlet distribution $D(\alpha_n)$, $\alpha_n = [\alpha_{n1}, \alpha_{n2}, \ldots, \alpha_{nm}]$, $\alpha_{nl} = \alpha_{0l} + \sum_{i=1}^{N} y_{il}$, $l = 1, 2, \ldots, m$. Observe that

$$\begin{aligned}\mathbb{E}[\theta_j|\mathbf{y}] &= \frac{\alpha_{nj}}{\sum_{l=1}^{m}\left[\alpha_{0l} + \sum_{i=1}^{N} y_{il}\right]} \\ &= \frac{\sum_{l=1}^{m} \alpha_{0l}}{\sum_{l=1}^{m}\left[\alpha_{0l} + \sum_{i=1}^{N} y_{il}\right]} \frac{\alpha_{0j}}{\sum_{l=1}^{m} \alpha_{0l}} \\ &+ \frac{\sum_{l=1}^{m}\sum_{i=1}^{N} y_{il}}{\sum_{l=1}^{m}\left[\alpha_{0l} + \sum_{i=1}^{N} y_{il}\right]} \frac{\sum_{i=1}^{N} y_{ij}}{\sum_{l=1}^{m}\sum_{i=1}^{N} y_{il}}.\end{aligned}$$

We have again that the posterior mean is a weighted average between the prior mean and the maximum likelihood estimate.

The marginal likelihood is

$$\begin{aligned} p(\mathbf{y}) &= \int_{\mathbf{\Theta}} \frac{\prod_{l=1}^{m} \theta_l^{\alpha_{0l}-1}}{B(\alpha_0)} \times \prod_{i=1}^{N} \frac{n!}{\prod_{l=1}^{m} y_{il}} \prod_{l=1}^{m} \theta_l^{y_{il}} d\theta \\ &= \frac{N \times n!}{B(\alpha_0) \prod_{i=1}^{N} \prod_{l=1}^{m} y_{il}!} \int_{\mathbf{\Theta}} \prod_{l=1}^{m} \theta_l^{\alpha_{0l}+\sum_{i=1}^{N} y_{il}-1} d\theta \\ &= \frac{N \times n!}{B(\alpha_0) \prod_{i=1}^{N} \prod_{l=1}^{m} y_{il}!} B(\alpha_n) \\ &= \frac{N \times n! \Gamma\left(\sum_{l=1}^{n} \alpha_{0l}\right)}{\Gamma\left(\sum_{l=1}^{n} \alpha_{0l} + N \times n\right)} \prod_{l=1}^{m} \frac{\Gamma\left(\alpha_{nl}\right)}{\Gamma\left(\alpha_{0l}\right) \prod_{i=1}^{N} y_{il}!}, \end{aligned}$$

where $B(\alpha) = \frac{\prod_{l=1}^{m} \Gamma(\alpha_l)}{\Gamma\left(\sum_{l=1}^{m} \alpha_l\right)}$.

Following similar steps we get the predictive density

$$p(Y_0|\mathbf{y}) = \frac{n! \Gamma\left(\sum_{l=1}^{n} \alpha_{nl}\right)}{\Gamma\left(\sum_{l=1}^{n} \alpha_{nl} + n\right)} \prod_{l=1}^{m} \frac{\Gamma\left(\alpha_{nl} + y_{0l}\right)}{\Gamma\left(\alpha_{nl}\right) y_{0l}!}.$$

This is a Dirichlet-multinomial distribution with parameters α_n.

2. **Likelihood functions from continuous distributions**

   The normal-normal/inverse-gamma model

   Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ from a normal distribution, then the conjugate prior density has the form

$$
\begin{aligned}
\pi(\mu,\sigma^2) &\propto \exp\left\{b_0\left(-\frac{\mu^2}{2\sigma^2}-\frac{\log\sigma^2}{2}\right)\right\}\exp\left\{a_{01}\frac{\mu}{\sigma^2}-a_{02}\frac{1}{\sigma^2}\right\}\\
&= \exp\left\{b_0\left(-\frac{\mu^2}{2\sigma^2}-\frac{\log\sigma^2}{2}\right)\right\}\exp\left\{a_{01}\frac{\mu}{\sigma^2}-a_{02}\frac{1}{\sigma^2}\right\}\\
&\times \exp\left\{-\frac{a_{01}^2}{2\sigma^2 b_0}\right\}\exp\left\{\frac{a_{01}^2}{2\sigma^2 b_0}\right\}\\
&= \exp\left\{-\frac{b_0}{2\sigma^2}\left(\mu-\frac{a_{01}}{b_0}\right)^2\right\}\left(\frac{1}{\sigma^2}\right)^{\frac{b_0+1-1}{2}}\\
&\times \exp\left\{\frac{1}{\sigma^2}\frac{-2b_0a_{02}+a_{01}^2}{2b_0}\right\}\\
&= \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{b_0}{2\sigma^2}\left(\mu-\frac{a_{01}}{b_0}\right)^2\right\}}_{1}\\
&\times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{b_0-1}{2}}\exp\left\{-\frac{1}{\sigma^2}\frac{2b_0a_{02}-a_{01}^2}{2b_0}\right\}}_{2}.
\end{aligned}
$$

The first part is the kernel of a normal density with mean $\mu_0 = a_{01}/\beta_0$ and variance σ^2/β_0, $\beta_0 = b_0$ that is, $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\beta_0)$. The second part is the kernel of an inverse gamma density with shape parameter $\alpha_0/2 = \frac{\beta_0-3}{2}$, and scale parameter $\delta_0/2 = \frac{2\beta_0 a_{02}-a_{01}^2}{2\beta_0}$, $\sigma^2 \sim IG(\alpha_0/2, \delta_0/2)$. Observe that $b_0 = \beta_0$ is the hypothetical sample size, and a_{01} is the hypothetical sum of prior observations, then, it makes sense that a_{01}/β_0 and σ^2/β_0 are the prior mean and variance, respectively.

Therefore, the posterior distribution is also a normal-inverse gamma

distribution,

$$\begin{aligned}
\pi(\mu,\sigma^2|\mathbf{y}) &\propto \left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{\beta_0}{2\sigma^2}(\mu-\mu_0)^2\right\}\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\exp\left\{-\frac{\delta_0}{2\sigma^2}\right\}\\
&\quad\times(\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}\\
&= \left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\beta_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(y_i-\bar{y})^2+N(\mu-\bar{y})^2+\delta_0\right)\right\}\\
&\quad\times\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N}{2}+1}+\frac{(\beta_0\mu_0+N\bar{y})^2}{\beta_0+N}-\frac{(\beta_0\mu_0+N\bar{y})^2}{\beta_0+N}\\
&= \underbrace{\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left((\beta_0+N)\left(\mu-\left(\frac{\beta_0\mu_0+N\bar{y}}{\beta_0+N}\right)\right)^2\right)\right\}}_{1}\\
&\quad\times\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(y_i-\bar{y})^2+\delta_0+\frac{\beta_0 N}{\beta_0+N}(\bar{y}-\mu_0)^2\right)\right\}}_{2}.
\end{aligned}$$

The first term is the kernel of a normal density, $\mu|\sigma^2,\mathbf{y}\sim N\left(\mu_n,\sigma_n^2\right)$, where $\mu_n=\frac{\beta_0\mu_0+N\bar{y}}{\beta_0+N}$ and $\sigma_n^2=\frac{\sigma^2}{\beta_n}$, $\beta_n=\beta_0+N$. The second term is the kernel of an inverse gamma density, $\sigma^2|\mathbf{y}\sim IG(\alpha_n/2,\delta_n/2)$ where $\alpha_n=\alpha_0+N$ and $\delta_n=\sum_{i=1}^{N}(y_i-\bar{y})^2+\delta_0+\frac{\beta_0 N}{\beta_0+N}(\bar{y}-\mu_0)^2$. Observe that the posterior mean is a weighted average between prior and sample information. The weights depends on the sample sizes (β_0 and N).

The marginal posterior for σ^2 is inverse gamma with shape and scale parameters $\alpha_n/2$ and $\delta_n/2$, respectively. The marginal posterior of μ is

$$\begin{aligned}
\pi(\mu|\mathbf{y}) &\propto \int_0^\infty\left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+1}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}(\beta_n(\mu-\mu_n)^2+\delta_n)\right\}\right\}d\sigma^2\\
&= \frac{\Gamma\left(\frac{\alpha_n+1}{2}\right)}{\left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{\frac{\alpha_n+1}{2}}}\\
&\propto \left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{-\frac{\alpha_n+1}{2}}\left(\frac{\delta_n}{\delta_n}\right)^{-\frac{\alpha_n+1}{2}}\\
&\propto \left[\frac{\alpha_n\beta_n(\mu-\mu_n)^2}{\alpha_n\delta_n}+1\right]^{-\frac{\alpha_n+1}{2}},
\end{aligned}$$

where the second line due to having the kernel of an inverse gamma density with parameters $(\alpha_n+1)/2$ and $-\frac{1}{2\sigma^2}(\beta_n(\mu-\mu_n)^2+\delta_n)$.

This is the kernel of a Student's t distribution, $\mu|\mathbf{y} \sim t(\mu_n, \delta_n/\beta_n\alpha_n, \alpha_n)$, where $\mathbb{E}[\mu|\mathbf{y}] = \mu_n$ and $Var[\mu|\mathbf{y}] = \frac{\alpha_n}{\alpha_n-2}\left(\frac{\delta_n}{\beta_n\alpha_n}\right) = \frac{\delta_n}{(\alpha_n-2)\beta_n}$, $\alpha_n > 2$. Observe that the marginal posterior distribution for μ has heavier tails than the conditional posterior distribution due to incorporating uncertainty regarding σ^2. The marginal likelihood is

$$
\begin{aligned}
p(\mathbf{y}) &= \int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{(2\pi\sigma^2/\beta_0)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2/\beta_0}(\mu-\mu_0)^2\right\}\frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\right. \\
&\left.\times\exp\left\{-\frac{\delta_0}{2\sigma^2}\right\}(2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}\right\}d\sigma^2 d\mu \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N+1}{2}+1}\right. \\
&\left.\times\exp\left\{-\frac{1}{2\sigma^2}(\beta_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(y_i-\mu)^2+\delta_0)\right\}\right\}d\sigma^2 d\mu \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\Gamma\left(\frac{N+1+\alpha_0}{2}\right) \\
&\times\int_{-\infty}^{\infty}\left[\frac{\beta_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(y_i-\mu)^2+\delta_0}{2}\right]^{-\frac{\alpha_0+N+1}{2}}d\mu \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\Gamma\left(\frac{N+1+\alpha_0}{2}\right) \\
&\times\int_{-\infty}^{\infty}\left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{-\frac{\alpha_n+1}{2}}d\mu\left(\frac{\delta_n/2}{\delta_n/2}\right)^{-\frac{\alpha_n+1}{2}} \\
&= \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}(2\pi)^{-\left(\frac{N+1}{2}\right)}\beta_0^{1/2}\Gamma\left(\frac{\alpha_n+1}{2}\right)\left(\frac{\delta_n}{2}\right)^{-\frac{\alpha_n+1}{2}}\frac{\left(\frac{\delta_n\pi}{\beta_n}\right)^{1/2}\Gamma\left(\frac{\alpha_n}{2}\right)}{\Gamma\left(\frac{\alpha_n+1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\alpha_n}{2}\right)}{\Gamma\left(\frac{\alpha_0}{2}\right)}\frac{(\delta_0/2)^{\alpha_0/2}}{(\delta_n/2)^{\alpha_n/2}}\left(\frac{\beta_0}{\beta_n}\right)^{1/2}(\pi)^{-N/2},
\end{aligned}
$$

where we take into account that $\int_{-\infty}^{\infty}\left[\frac{\beta_n(\mu-\mu_n)^2+\delta_n}{2}\right]^{-\frac{\alpha_n+1}{2}}d\mu\left(\frac{\delta_n/2}{\delta_n/2}\right)^{-\frac{\alpha_n+1}{2}} = \int_{-\infty}^{\infty}\left[\frac{\beta_n\alpha_n(\mu-\mu_n)^2}{\delta_n\alpha_n}+1\right]^{-\frac{\alpha_n+1}{2}}d\mu\left(\frac{\delta_n}{2}\right)^{-\frac{\alpha_n+1}{2}}$. The term in the integral is the kernel of a Student's t density, this means that the integral is equal to $\frac{\left(\frac{\delta_n\pi}{\beta_n}\right)^{1/2}\Gamma\left(\frac{\alpha_n}{2}\right)}{\Gamma\left(\frac{\alpha_n+1}{2}\right)}$.

The predictive density is

$$\begin{aligned}
\pi(Y_0|\mathbf{y}) &\propto \int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(y_0-\mu)^2\right\}\left(\frac{1}{\sigma^2}\right)^{1/2}\exp\left\{-\frac{\beta_n}{2\sigma^2}(\mu-\mu_n)^2\right\}\right. \\
&\quad\left.\times\left(\frac{1}{\sigma^2}\right)^{\alpha_n/2+1}\exp\left\{-\frac{\delta_n}{2\sigma^2}\right\}\right\}d\sigma^2 d\mu \\
&= \int_{-\infty}^{\infty}\int_{0}^{\infty}\left\{\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+2}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}((y_0-\mu)^2+\beta_n(\mu-\mu_n)^2+\delta_n)\right\}\right\}d\sigma^2 d\mu \\
&\propto \int_{-\infty}^{\infty}\left[\beta_n(\mu-\mu_n)^2+(y_0-\mu)^2+\delta_n\right]^{-\left(\frac{\alpha_n}{2}+1\right)}d\mu \\
&= \int_{-\infty}^{\infty}\left[(\beta_n+1)\left(\mu-\left(\frac{\beta_n\mu_n+y_0}{\beta_n+1}\right)\right)^2+\frac{\beta_n(y_0-\mu_n)^2}{\beta_n+1}+\delta_n\right]^{-\left(\frac{\alpha_n}{2}+1\right)}d\mu \\
&= \int_{-\infty}^{\infty}\left[1+\frac{(\beta_n+1)^2\left(\mu-\left(\frac{\beta_n\mu_n+y_0}{\beta_n+1}\right)\right)^2}{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}\right]^{-\left(\frac{\alpha_n}{2}+1\right)}d\mu \\
&\quad\times\left(\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{\beta_n+1}\right)^{-\left(\frac{\alpha_n}{2}+1\right)} \\
&\propto \left(\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{(\beta_n+1)^2(\alpha_n+1)}\right)^{\frac{1}{2}}\left(\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{\beta_n+1}\right)^{-\left(\frac{\alpha_n}{2}+1\right)} \\
&\propto (\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n)^{\left(\frac{\alpha_n+1}{2}\right)} \\
&\propto \left[1+\frac{\beta_n\alpha_n}{(\beta_n+1)\delta_n\alpha_n}(y_0-\mu_n)^2\right]^{-\left(\frac{\alpha_n+1}{2}\right)},
\end{aligned}$$

where we have that $\left[1+\frac{(\beta_n+1)^2\left(\mu-\left(\frac{\beta_n\mu_n+y_0}{\beta_n+1}\right)\right)^2}{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}\right]^{-\left(\frac{\alpha_n}{2}+1\right)}$ is the kernel of a Student's t density with degrees of freedom α_n+1 and scale $\frac{\beta_n(y_0-\mu_n)^2+(\beta_n+1)\delta_n}{(\beta_n+1)^2(\alpha_n+1)}$.

The last expression is the kernel of a Student's t density, that is, $Y_0|\mathbf{y}\sim t\left(\mu_n,\frac{(\beta_n+1)\delta_n}{\beta_n\alpha_n},\alpha_n\right)$.

The multivariate normal-normal/inverse-Wishart model

We show in subsection 4.1 that the multivariate normal distribution is in the exponential family where

$$C(\mu,\mathbf{\Sigma}) = \exp\left\{-\frac{1}{2}\left(tr\left(\mu\mu^\top\mathbf{\Sigma}^{-1}\right)+\log(|\Sigma|)\right)\right\},$$

$$\eta(\mu,\mathbf{\Sigma})^\top = \left[\left(vec\left(\mathbf{\Sigma}^{-1}\right)\right)^\top \quad \left(vec\left(\mu^\top\mathbf{\Sigma}^{-1}\right)\right)^\top\right],$$

$$T(\mathbf{y}) = \left[-\frac{1}{2}\left(vec\left(\mathbf{S}\right)^\top+Nvec\left(\hat{\mu}\hat{\mu}^\top\right)^\top\right) \quad -N\hat{\mu}^\top\right]^\top$$

and

$$h(\mathbf{y}) = (2\pi)^{-pN/2}.$$

Then, its conjugate prior distribution should have the form

$$\begin{aligned}
\pi(\mu, \mathbf{\Sigma}) &\propto \exp\left\{-\frac{b_0}{2}\left(tr\left(\mu\mu^\top \mathbf{\Sigma}^{-1}\right) + \log(|\Sigma|)\right)\right\} \\
&\times \exp\left\{\mathbf{a}_{01}^\top vec\left(\mathbf{\Sigma}^{-1}\right) + \mathbf{a}_{02}^\top vec\left(\mu^\top \mathbf{\Sigma}^{-1}\right)\right\} \\
&= |\Sigma|^{-b_0/2} \exp\left\{-\frac{b_0}{2}\left(tr\left(\mu^\top \mathbf{\Sigma}^{-1}\mu\right)\right) + tr\left(\mathbf{a}_{02}^\top \mathbf{\Sigma}^{-1}\mu\right)\right\} \\
&\times \exp\left\{\mathbf{a}_{01}^\top vec\left(\mathbf{\Sigma}^{-1}\right) + \frac{\mathbf{a}_{02}^\top \mathbf{\Sigma}^{-1}\mathbf{a}_{02}}{2b_0} - \frac{\mathbf{a}_{02}^\top \mathbf{\Sigma}^{-1}\mathbf{a}_{02}}{2b_0}\right\} \\
&= |\Sigma|^{-b_0/2} \exp\left\{-\frac{b_0}{2}\left(\mu - \frac{\mathbf{a}_{02}}{b_0}\right)^\top \mathbf{\Sigma}^{-1}\left(\mu - \frac{\mathbf{a}_{02}}{b_0}\right)\right\} \\
&\times \exp\left\{-\frac{1}{2}tr\left(\left(\mathbf{A}_{01} - \frac{\mathbf{a}_{02}\mathbf{a}_{02}^\top}{b_0}\right)\mathbf{\Sigma}^{-1}\right)\right\} \\
&= \underbrace{|\Sigma|^{-1/2} \exp\left\{-\frac{b_0}{2}\left(\mu - \frac{\mathbf{a}_{02}}{b_0}\right)^\top \mathbf{\Sigma}^{-1}\left(\mu - \frac{\mathbf{a}_{02}}{b_0}\right)\right\}}_{1} \\
&\times \underbrace{|\Sigma|^{-(\alpha_0+p+1)/2} \exp\left\{-\frac{1}{2}tr\left(\left(\mathbf{A}_{01} - \frac{\mathbf{a}_{02}\mathbf{a}_{02}^\top}{b_0}\right)\mathbf{\Sigma}^{-1}\right)\right\}}_{2},
\end{aligned}$$

where b_0 is the hypothetical sample size, and $\mathbf{a}_{01}$ and $\mathbf{a}_{02}$ are p^2 and p dimensional vectors of prior sufficient statistics, where $\mathbf{a}_{01} = -\frac{1}{2}vec(\mathbf{A}_{01})$ such that $\mathbf{A}_{01}$ is a $p \times p$ positive semi-definite matrix. Setting $b_0 = 1 + \alpha_0 + p + 1$ we have that the first part in the last expression is the kernel of a multivariate normal density with mean $\mu_0 = \mathbf{a}_{02}/b_0$ and covariance $\frac{\mathbf{\Sigma}}{b_0}$, that is, $\mu|\mathbf{\Sigma} \sim N_p\left(\mu_0, \frac{\mathbf{\Sigma}}{\beta_0}\right)$, $b_0 = \beta_0$. It makes sense these hyperparameters because $\mathbf{a}_{02}$ is the hypothetical sum of prior observations and b_0 is the hypothetical prior sample size. On the other hand, the second expression in the last line is the kernel of a Inverse-Wishart distribution with scale matrix $\mathbf{\Psi}_0 = \left(\mathbf{A}_{01} - \frac{\mathbf{a}_{02}\mathbf{a}_{02}^\top}{b_0}\right)$ and degrees of freedom α_0, that is, $\mathbf{\Sigma} \sim IW_p(\mathbf{\Psi}_0, \alpha_0)$. Observe that $\mathbf{\Psi}_0$ has the same structure as the first part of the sufficient statistics in $T(\mathbf{y})$, just that it should be understood as coming from prior hypothetical observations.

Therefore, the prior distribution in this setting is normal/inverse-Wishart, and given conjugacy, the posterior distribution is in the

same family.

$$\pi(\mu,\boldsymbol{\Sigma}|\mathbf{Y}) \propto (2\pi)^{-pN/2}|\Sigma|^{-N/2}\exp\left\{-\frac{1}{2}tr\left[\left(\mathbf{S}+N\left(\mu-\hat{\mu}\right)\left(\mu-\hat{\mu}\right)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}$$
$$\times|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{\beta_0}{2}tr\left[(\mu-\mu_0)(\mu-\mu_0)^{\top}\boldsymbol{\Sigma}^{-1}\right]\right\}|\boldsymbol{\Sigma}|^{-(\alpha_0+p+1)/2}$$
$$\times\exp\left\{-\frac{1}{2}tr(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}^{-1})\right\}.$$

Taking into account that

$$N\left(\mu-\hat{\mu}\right)\left(\mu-\hat{\mu}\right)^{\top}+\beta_0\left(\mu-\mu_0\right)\left(\mu-\mu_0\right)^{\top}=\left(N+\beta_0\right)\left(\mu-\mu_n\right)\left(\mu-\mu_n\right)^{\top}$$
$$+\frac{N\beta_0}{N+\beta_0}\left(\hat{\mu}-\mu_0\right)\left(\hat{\mu}-\mu_0\right)^{\top},$$

where $\mu_n=\frac{N}{N+\beta_0}\hat{\mu}+\frac{\beta_0}{N+\beta_0}\mu_0$ is the posterior mean. We have

$$\pi(\mu,\boldsymbol{\Sigma}|\mathbf{Y}) \propto |\Sigma|^{-1/2}\exp\left\{-\frac{N+\beta_0}{2}tr\left[\left(\left(\mu-\mu_n\right)\left(\mu-\mu_n\right)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}$$
$$\times|\boldsymbol{\Sigma}|^{-(N+\alpha_0+p+1)/2}$$
$$\times\exp\left\{-\frac{1}{2}tr\left[\left(\boldsymbol{\Psi}_0+\mathbf{S}+\frac{N\beta_0}{N+\beta_0}(\hat{\mu}-\mu_0)(\hat{\mu}-\mu_0)^{\top}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

Then, $\mu|\boldsymbol{\Sigma},\mathbf{Y}\sim N_p\left(\mu_n,\frac{1}{\beta_n}\boldsymbol{\Sigma}\right)$, and $\boldsymbol{\Sigma}|\mathbf{Y}\sim IW\left(\boldsymbol{\Psi}_n,\alpha_n\right)$ where $\beta_n=N+\beta_0$, $\alpha_n=N+\alpha_0$ and $\boldsymbol{\Psi}_n=\boldsymbol{\Psi}_0+\mathbf{S}+\frac{N\beta_0}{N+\beta_0}(\hat{\mu}-\mu_0)(\hat{\mu}-\mu_0)^{\top}$.

The marginal posterior of μ is given by $\int_{\mathcal{S}}\pi(\mu,\boldsymbol{\Sigma})d\boldsymbol{\Sigma}$ where $\mathcal{S}$ is the space of positive semi-definite matrices. Then,

$$\pi(\mu|\mathbf{Y}) \propto \int_{\mathcal{S}}\left\{|\boldsymbol{\Sigma}|^{-(\alpha_n+p+2)/2}\right.$$
$$\left.\exp\left\{-\frac{1}{2}tr\left[\left(\beta_n\left(\mu-\mu_n\right)\left(\mu-\mu_n\right)^{\top}+\boldsymbol{\Psi}_n\right)\boldsymbol{\Sigma}^{-1}\right]\right\}\right\}d\boldsymbol{\Sigma}$$
$$\propto|\left(\beta_n\left(\mu-\mu_n\right)\left(\mu-\mu_n\right)^{\top}+\boldsymbol{\Psi}_n\right)|^{-(\alpha_n+1)/2}$$
$$=\left[\left|\boldsymbol{\Psi}_n\right|\times\left|1+\beta_n\left(\mu-\mu_n\right)^{\top}\boldsymbol{\Psi}_n^{-1}\left(\mu-\mu_n\right)\right|\right]^{-(\alpha_n+1)/2}$$
$$\propto\left(1+\frac{1}{\alpha_n+1-p}\left(\mu-\mu_n\right)^{\top}\left(\frac{\boldsymbol{\Psi}_n}{(\alpha_n+1-p)\beta_n}\right)^{-1}\left(\mu-\mu_n\right)\right)^{-(\alpha_n+1-p+p)/2},$$

where the second line uses properties of the inverse Wishart distribution, and the third line uses a particular case of the Sylvester's determinant theorem.

We observe that the last line is the kernel of a Multivariate Student's t distribution, that is, $\mu|\mathbf{Y}\sim t_p(v_n,\mu_n,\boldsymbol{\Sigma}_n)$ where $v_n=\alpha_n+1-p$ and $\boldsymbol{\Sigma}_n=\frac{\boldsymbol{\Psi}_n}{(\alpha_n+1-p)\beta_n}$.

The marginal likelihood is given by

$$p(\mathbf{Y})=\frac{\Gamma_p\left(\frac{v_n}{2}\right)}{\Gamma_p\left(\frac{\alpha_0}{2}\right)}\frac{|\boldsymbol{\Psi}_0|^{\alpha_0/2}}{|\boldsymbol{\Psi}_n|^{\alpha_n/2}}\left(\frac{\beta_0}{\beta_n}\right)^{p/2}(2\pi)^{-Np/2},$$

where Γ_p is the multivariate gamma function (see Exercise 5).

The posterior predictive distribution is $\mathbf{Y}_0|\mathbf{Y} \sim t_p(v_n, \mu_n, (\beta_n + 1)\mathbf{\Sigma}_n)$ (see Exercise 6).

4.3 Linear regression: The conjugate normal-normal/inverse gamma model

In this setting we analyze the conjugate normal-normal/inverse gamma model which is the workhorse in econometrics. In this model, the dependent variable y_i is related to a set of regressors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})^\top$ in a linear way, that is, $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \mu_i = \mathbf{x}_i^\top \beta + \mu_i$ where $\beta = (\beta_1, \beta_2, \ldots, \beta_K)^\top$ and $\mu_i \overset{iid}{\sim} N(0, \sigma^2)$ is an stochastic error such that $\mathbb{E}[\mu_i|\mathbf{x}_i] = 0$.

Defining $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1K} \\ x_{21} & x_{22} & \ldots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{NK} \end{bmatrix}$ and $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$,

we can write the model in matrix form: $\mathbf{y} = \mathbf{X}\beta + \mu$, where $\mu \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ which implies that $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Then, the likelihood function is

$$\begin{aligned} p(\mathbf{y}|\beta, \sigma^2, \mathbf{X}) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\right\} \\ &\propto (\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\right\}. \end{aligned}$$

The conjugate priors for the parameters are

$$\begin{aligned} \beta|\sigma^2 &\sim N(\beta_0, \sigma^2\mathbf{B}_0), \\ \sigma^2 &\sim IG(\alpha_0/2, \delta_0/2). \end{aligned}$$

Then, the posterior distribution is

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

where $\hat{\beta}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the maximum likelihood estimator.

Adding and subtracting $\beta_n^\top\mathbf{B}_n^{-1}\beta_n$ to complete the square, where $\mathbf{B}_n=(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}$ and $\beta_n=\mathbf{B}_n(\mathbf{B}_0^{-1}\beta_0+\mathbf{X}^\top\mathbf{X}\hat{\beta})$,

$$\begin{aligned}\pi(\beta,\sigma^2|\mathbf{y},\mathbf{X}) \propto &\underbrace{(\sigma^2)^{-\frac{K}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\beta_n)^\top\mathbf{B}_n^{-1}(\beta-\beta_n)\right\}}_{1}\\ &\times\underbrace{(\sigma^2)^{-\left(\frac{\alpha_n}{2}+1\right)}\exp\left\{-\frac{\delta_n}{2\sigma^2}\right\}}_{2}.\end{aligned}$$

The first expression is the kernel of a normal density function, $\beta|\sigma^2,\mathbf{y},\mathbf{X}\sim N(\beta_n,\sigma^2\mathbf{B}_n)$. The second expression is the kernel of a inverse gamma density, $\sigma^2|\mathbf{y},\mathbf{X}\sim IG(\alpha_n/2,\delta_n/2)$, where $\alpha_n=\alpha_0+N$ and $\delta_n=\delta_0+\mathbf{y}^\top\mathbf{y}+\beta_0^\top\mathbf{B}_0^{-1}\beta_0-\beta_n^\top\mathbf{B}_n^{-1}\beta_n$.

Taking into account that

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

where $(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{B}_0^{-1}=\mathbf{I_K}-(\mathbf{B_0^{-1}}+\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}$ [59]. Setting $\mathbf{W}=(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}$ we have $\beta_n=(\mathbf{I_K}-\mathbf{W})\beta_\mathbf{0}+\mathbf{W}\hat{\beta}$, that is, the posterior mean of β is a weighted average between the sample and prior information, where the weights depend on the precision of each piece of information. Observe that when the prior covariance matrix is highly vague (non–informative), such that $\mathbf{B}_0^{-1}\to\mathbf{0_K}$, we obtain $\mathbf{W}\to I_K$, such that $\beta_n\to\hat{\beta}$, that is, the posterior mean location parameter converges to the maximum likelihood estimator.

In addition, we know that the posterior conditional covariance matrix of the location parameters $\sigma^2(\mathbf{B}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}=\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}-$

$\sigma^2\left((\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{B}_0+(\mathbf{X}^\top\mathbf{X})^{-1})^{-1}(\mathbf{X}^\top\mathbf{X})^{-1}\right)$ is positive semi-definite.[2] Given that $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is the covariance matrix of the maximum likelihood estimator, we observe that prior information reduces estimation uncertainty.

Now, we calculate the posterior marginal distribution of β,

$$\begin{aligned}\pi(\beta|\mathbf{y},\mathbf{X}) &= \int_0^\infty \pi(\beta,\sigma^2|\mathbf{y},\mathbf{X})d\sigma^2\\ &= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+K}{2}+1}\exp\left\{-\frac{s}{2\sigma^2}\right\}d\sigma^2,\end{aligned}$$

where $s=\delta_n+(\beta-\beta_n)^\top\mathbf{B}_n^{-1}(\beta-\beta_n)$. Then we can write

$$\begin{aligned}\pi(\beta|\mathbf{y},\mathbf{X}) &= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_n+K}{2}+1}\exp\left\{-\frac{s}{2\sigma^2}\right\}d\sigma^2\\ &= \frac{\Gamma((\alpha_n+K)/2)}{(s/2)^{(\alpha_n+K)/2}}\int_0^\infty \frac{(s/2)^{(\alpha_n+K)/2}}{\Gamma((\alpha_n+K)/2)}(\sigma^2)^{-(\alpha_n+K)/2-1}\exp\left\{-\frac{s}{2\sigma^2}\right\}d\sigma^2.\end{aligned}$$

The right term is the integral of the probability density function of an inverse gamma distribution with parameters $\nu=(\alpha_n+K)/2$ and $\tau=s/2$. Since we are integrating over the whole support of σ^2, the integral is equal to 1, and therefore

$$\begin{aligned}\pi(\beta|\mathbf{y},\mathbf{X}) &= \frac{\Gamma((\alpha_n+K)/2)}{(s/2)^{(\alpha_n+K)/2}}\\ &\propto s^{-(\alpha_n+K)/2}\\ &= [\delta_n+(\beta-\beta_n)^\top\mathbf{B}_n^{-1}(\beta-\beta_n)]^{-(\alpha_n+K)/2}\\ &= \left[1+\frac{(\beta-\beta_n)^\top\left(\frac{\delta_n}{\alpha_n}\mathbf{B}_n\right)^{-1}(\beta-\beta_n)}{\alpha_n}\right]^{-(\alpha_n+K)/2}(\delta_n)^{-(\alpha_N+K)/2}\\ &\propto \left[1+\frac{(\beta-\beta_n)^\top\mathbf{H}_n^{-1}(\beta-\beta_n)}{\alpha_n}\right]^{-(\alpha_n+K)/2},\end{aligned}$$

where $\mathbf{H}_n=\frac{\delta_n}{\alpha_n}\mathbf{B}_n$. This last expression is a multivariate Student's t distribution for β, $\beta|\mathbf{y},\mathbf{X}\sim t_K(\alpha_n,\beta_n,\mathbf{H}_n)$.

Observe that as we have incorporated the uncertainty of the variance, the posterior for β changes from a normal to a Students' t distribution, which has heavier tails, indicating more uncertainty.

The marginal likelihood of this model is

$$p(\mathbf{y})=\int_0^\infty\int_{R^K}\pi(\beta|\sigma^2,\mathbf{B}_0,\beta_0)\pi(\sigma^2|\alpha_0/2,\delta_0/2)p(\mathbf{y}|\beta,\sigma^2,\mathbf{X})d\sigma^2d\beta.$$

[2] A particular case of the Woodbury matrix identity

Taking into account that $(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)+(\beta-\beta_0)^\top\mathbf{B}_0^{-1}(\beta-\beta_0)=(\beta-\beta_n)^\top\mathbf{B}_n^{-1}(\beta-\beta_n)+m$, where $m=\mathbf{y}^\top\mathbf{y}+\beta_0^\top\mathbf{B}_0^{-1}\beta_0-\beta_n^\top\mathbf{B}_n^{-1}\beta_n$, we have that

$$
\begin{aligned}
p(\mathbf{y}) &= \int_0^\infty \int_{R^K} \pi(\beta|\sigma^2)\pi(\sigma^2)p(\mathbf{y}|\beta,\sigma^2,\mathbf{X})d\sigma^2 d\beta \\
&= \int_0^\infty \pi(\sigma^2)\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}m\right\}\frac{1}{(2\pi\sigma^2)^{K/2}|\mathbf{B}_0|^{1/2}} \\
&\times \int_{R^K}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\beta_n)^\top\mathbf{B}_n^{-1}(\beta-\beta_n)\right\}d\sigma^2 d\beta \\
&= \int_0^\infty \pi(\sigma^2)\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}m\right\}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}d\sigma^2 \\
&= \int_0^\infty \frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1}\exp\left\{\left(-\frac{\delta_0}{2\sigma^2}\right)\right\}\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}m\right\}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}d\sigma^2 \\
&= \frac{1}{(2\pi)^{N/2}}\frac{(\delta_0/2)^{\alpha_0/2}}{\Gamma(\alpha_0/2)}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}\int_0^\infty\left(\frac{1}{\sigma^2}\right)^{\frac{\alpha_0+N}{2}+1}\exp\left\{\left(-\frac{\delta_0+m}{2\sigma^2}\right)\right\}d\sigma^2 \\
&= \frac{1}{\pi^{N/2}}\frac{\delta_0^{\alpha_0/2}}{\delta_n^{\alpha_n/2}}\frac{|\mathbf{B}_n|^{1/2}}{|\mathbf{B}_0|^{1/2}}\frac{\Gamma(\alpha_n/2)}{\Gamma(\alpha_0/2)}.
\end{aligned}
$$

We can show that $\delta_n=\delta_0+\mathbf{y}^\top\mathbf{y}+\beta_0^\top\mathbf{B}_0^{-1}\beta_0-\beta_n^\top\mathbf{B}_n^{-1}\beta_n=\delta_0+(\mathbf{y}-\mathbf{X}\hat{\beta})^\top(\mathbf{y}-\mathbf{X}\hat{\beta})+(\hat{\beta}-\beta_0)^\top((\mathbf{X}^\top\mathbf{X})^{-1}+\mathbf{B}_0)^{-1}(\hat{\beta}-\beta_0)$ (see Exercise 7). Therefore, if we want to compare two models under this setting, the Bayes factor is

$$
\begin{aligned}
BF_{12} &= \frac{p(\mathbf{y}|\mathcal{M}_\mathbf{1})}{p(\mathbf{y}|\mathcal{M}_\mathbf{2})} \\
&= \frac{\frac{\delta_{10}^{\alpha_{10}/2}}{\delta_{1n}^{\alpha_{1n}/2}}\frac{|\mathbf{B}_{1n}|^{1/2}}{|\mathbf{B}_{10}|^{1/2}}\frac{\Gamma(\alpha_{1n}/2)}{\Gamma(\alpha_{10}/2)}}{\frac{\delta_{20}^{\alpha_{20}/2}}{\delta_{2n}^{\alpha_{2n}/2}}\frac{|\mathbf{B}_{2n}|^{1/2}}{|\mathbf{B}_{20}|^{1/2}}\frac{\Gamma(\alpha_{2n}/2)}{\Gamma(\alpha_{20}/2)}},
\end{aligned}
$$

where subscripts 1 and 2 refer to each model, respectively.

Observe that *ceteris paribus*, the model having better fit, coherence between sample and prior information regarding location parameters, higher prior to posterior precision and less parameters is favored by the Bayes factor. Observe that the Bayes factor rewards model fit as the sum of squared errors is in δ_n, the better fit (lower sum of squared errors), the better the Bayes factor. In addition, a weighted distance between sample and prior location parameters also appears in δ_n, the greater this distance, the worse is model support. The ratio of determinants between posterior and prior covariance matrices is also present, the higher this ratio, the better for the Bayes factor supporting a model due to information gains. To see the effect of model's parsimony, let's take the common situation in applications where $\mathbf{B}_{j0}=c\mathbf{I}_{K_j}$ then $|\mathbf{B}_{j0}|=c^{K_j}$ such that $\left(\frac{|\mathbf{B}_{20}|}{|\mathbf{B}_{10}|}\right)^{1/2}=\left(\frac{c^{K_2/2}}{c^{K_1/2}}\right)$, if $K_2/K_1>1$ and $c\to\infty$, the latter implying a non-informative prior, then $BF_{12}\to\infty$, this means infinite evidence

supporting the parsimonious model no matter what sample information says. Comparing models having the same number of regressors ($K_1 = K_2$) is not a safe ground as $|\mathbf{B}_0|$ depending on measure units of the regressors such that conclusions regarding model selection depending on this, which is not a nice property. This prevents against using non-informative priors when performing model selection in the Bayesian framework. Observe that this is not the case when $\alpha_0 \to 0$ and $\delta_0 \to 0$, which implies a non-informative prior for the variance parameter.[3] We observe here that $\Gamma(\alpha_{j0})$ cancels out, $\alpha_{jn} \to N$ and $\delta_{jn} \to (\mathbf{y} - \mathbf{X}_j\hat{\beta}_j)^\top(\mathbf{y} - \mathbf{X}_j\hat{\beta}_j) + (\hat{\beta}_j - \beta_{j0})^\top((\mathbf{X}_j^\top\mathbf{X}_j)^{-1} + \mathbf{B}_{j0})^{-1}(\hat{\beta}_j - \beta_{j0})$, therefore there is not effect. This is due to σ^2 being a common parameter in both models. In general, we can use non-informative priors for common parameters to all models, but we cannot use non-informative prior for non-common parameters when performing model selection using the Bayes factor.

The posterior predictive is equal to

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) &= \int_0^\infty \int_{R^K} p(\mathbf{Y}_0|\beta,\sigma^2,\mathbf{y})\pi(\beta|\sigma^2,\mathbf{y})\pi(\sigma^2|\mathbf{y})d\beta d\sigma^2 \\ &= \int_0^\infty \int_{R^K} p(\mathbf{Y}_0|\beta,\sigma^2)\pi(\beta|\sigma^2,\mathbf{y})\pi(\sigma^2|\mathbf{y})d\beta d\sigma^2,\end{aligned}$$

where we take into account independence between $\mathbf{Y}_0$ and $\mathbf{Y}$. Given $\mathbf{X}_0$, which is the $N_0 \times K$ matrix of regressors associated with $\mathbf{Y}_0$, Then,

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) = \int_0^\infty \int_{R^K} \Bigg\{ &(2\pi\sigma^2)^{-\frac{N_0}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}_0 - \mathbf{X}_0\beta)^\top(\mathbf{Y}_0 - \mathbf{X}_0\beta)^\top\right\} \\ &\times (2\pi\sigma^2)^{-\frac{K}{2}}|\mathbf{B}_n|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_n)^\top\mathbf{B}_n^{-1}(\beta - \beta_n)\right\} \\ &\times \frac{(\delta_n/2)^{\alpha_n/2}}{\Gamma(\alpha_n/2)}\left(\frac{1}{\sigma^2}\right)^{\alpha_n/2+1}\exp\left\{-\frac{\delta_n}{2\sigma^2}\right\}\Bigg\} d\beta d\sigma^2.\end{aligned}$$

Setting $\mathbf{M} = (\mathbf{X}_0^\top\mathbf{X}_0 + \mathbf{B}_n^{-1})$ and $\beta_* = \mathbf{M}^{-1}(\mathbf{B}_n^{-1}\beta_n + \mathbf{X}_0^\top\mathbf{Y}_0)$, we have

$$\begin{aligned}(\mathbf{Y}_0 - \mathbf{X}_0\beta)^\top(\mathbf{Y}_0 - \mathbf{X}_0\beta)^\top + (\beta - \beta_n)^\top\mathbf{B}_n^{-1}(\beta - \beta_n) &= (\beta - \beta_*)^\top\mathbf{M}(\beta - \beta_*) \\ &\quad + \beta_n^\top\mathbf{B}_n^{-1}\beta_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \beta_*^\top\mathbf{M}\beta_*\end{aligned}$$

Thus,

$$\begin{aligned}\pi(\mathbf{Y}_0|\mathbf{y}) \propto \int_0^\infty \Bigg\{ &\left(\frac{1}{\sigma^2}\right)^{-\frac{K+N_0+\alpha_n}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}(\beta_n^\top\mathbf{B}_n^{-1}\beta_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \beta_*^\top\mathbf{M}\beta_* + \delta_n)\right\} \\ &\times \int_{R^K}\exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_*)^\top\mathbf{M}(\beta - \beta_*)\right\}d\beta\Bigg\} d\sigma^2,\end{aligned}$$

[3] [27] prevents against this common practice.

where the term in the second integral is the kernel of a multivariate normal density with mean β_* and covariance matrix $\sigma^2\mathbf{M}^{-1}$. Then,

$$\pi(\mathbf{Y}_0|\mathbf{y}) \propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{N_0+\alpha_n}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}(\beta_n^\top\mathbf{B}_n^{-1}\beta_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \beta_*^\top\mathbf{M}\beta_* + \delta_n)\right\} d\sigma^2,$$

which is the kernel of an inverse gamma density. Thus,

$$\pi(\mathbf{Y}_0|\mathbf{y}) \propto \left[\frac{\beta_n^\top\mathbf{B}_n^{-1}\beta_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \beta_*^\top\mathbf{M}\beta_* + \delta_n}{2}\right]^{-\frac{\alpha_n+N_0}{2}}.$$

Setting $\mathbf{C}^{-1} = \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{B}_n\mathbf{X}_0^\top$ such that $\mathbf{C} = \mathbf{I}_{N_0} - \mathbf{X}_0(\mathbf{B}_n^{-1} + \mathbf{X}_0^\top\mathbf{X}_0)^{-1}\mathbf{X}_0^\top = \mathbf{I}_{N_0} - \mathbf{X}_0\mathbf{M}^{-1}\mathbf{X}_0^\top$,[4] and $\beta_{**} = \mathbf{C}^{-1}\mathbf{X}_0\mathbf{M}^{-1}\mathbf{B}_n^{-1}\beta_n$, then

$$\begin{aligned}
\beta_n^\top\mathbf{B}_n^{-1}\beta_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - \beta_*^\top\mathbf{M}\beta_* &= \beta_n^\top\mathbf{B}_n^{-1}\beta_n + \mathbf{Y}_0^\top\mathbf{Y}_0 - (\beta_n^\top\mathbf{B}_n^{-1} + \mathbf{Y}_0^\top\mathbf{X}_0)\mathbf{M}^{-1}(\mathbf{B}_n^{-1}\beta_n + \mathbf{X}_0^\top\mathbf{Y}_0) \\
&= \beta_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\beta_n + \mathbf{Y}_0^\top\mathbf{C}\mathbf{Y}_0 \\
&\quad - 2\mathbf{Y}_0^\top\mathbf{C}\mathbf{C}^{-1}\mathbf{X}_0\mathbf{M}^{-1}\mathbf{B}_n^{-1}\beta_n + \beta_{**}^\top\mathbf{C}\beta_{**} - \beta_{**}^\top\mathbf{C}\beta_{**} \\
&= \beta_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\beta_n + (\mathbf{Y}_0 - \beta_{**})^\top\mathbf{C}(\mathbf{Y}_0 - \beta_{**}) \\
&\quad - \beta_{**}^\top\mathbf{C}\beta_{**},
\end{aligned}$$

where $\beta_n^\top(\mathbf{B}_n^{-1} - \mathbf{B}_n^{-1}\mathbf{M}^{-1}\mathbf{B}_n^{-1})\beta_n = \beta_{**}^\top\mathbf{C}\beta_{**}$ and $\beta_{**} = \mathbf{X}_0\beta_n$ (see Exercise 8).

Then,

$$\begin{aligned}
\pi(\mathbf{Y}_0|\mathbf{y}) &\propto \left[\frac{(\mathbf{Y}_0 - \mathbf{X}_0\beta_n)^\top\mathbf{C}(\mathbf{Y}_0 - \mathbf{X}_0\beta_n) + \delta_n}{2}\right]^{-\frac{\alpha_n+N_0}{2}} \\
&\propto \left[\frac{(\mathbf{Y}_0 - \mathbf{X}_0\beta_n)^\top\left(\frac{\mathbf{C}\alpha_n}{\delta_n}\right)(\mathbf{Y}_0 - \mathbf{X}_0\beta_n)}{\alpha_n} + 1\right]^{-\frac{\alpha_n+N_0}{2}}.
\end{aligned}$$

The posterior predictive is a multivariate Student's t, $\mathbf{Y}_0|\mathbf{y} \sim t\left(\mathbf{X}_0\beta_n, \frac{\delta_n(\mathbf{I}_{N_0}+\mathbf{X}_0\mathbf{B}_n\mathbf{X}_0^\top)}{\alpha_n}, \alpha_n\right)$.

4.4 Multivariate linear regression: The conjugate normal-normal/inverse Wishart model

Let's study the multivariate regression setting where there are N-dimensional vectors $\mathbf{y}_m$, $m = 1, 2, \ldots, M$ such that $\mathbf{y}_m = \mathbf{X}\beta_m + \mu_m$, $\mathbf{X}$ is the set of

[4] Using $(\mathbf{A} + \mathbf{BDC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}^{-1} + \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1}$

common regressors, and μ_m is the N-dimensional vector of stochastic errors for each equation such that $\mathbf{U} = [\mu_1\ \mu_2\ \ldots\ \mu_M] \sim MN_{N,M}(\mathbf{0}, \mathbf{I}_N, \mathbf{\Sigma})$, that is, a matrix variate normal distribution where $\mathbf{\Sigma}$ is the covariance matrix of each i-th row of $\mathbf{U}$, $i = 1, 2, \ldots, N$, and we are assuming independence between the rows. Then, $vec(\mathbf{U}) \sim N_{N\times M}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I_N})$.[5]

This framework can be written in matrix form

$$\underbrace{\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ldots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KM} \end{bmatrix}}_{\mathbf{B}}$$
$$+ \underbrace{\begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1M} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ \mu_{N1} & \mu_{N2} & \cdots & \mu_{NM} \end{bmatrix}}_{\mathbf{U}}.$$

Therefore, $\mathbf{Y} \sim N_{N\times M}(\mathbf{XB}, \mathbf{\Sigma} \otimes \mathbf{I_N})$,[6]

$$\begin{aligned} p(\mathbf{Y}|\mathbf{B}, \mathbf{\Sigma}) &\propto |\mathbf{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2} tr\left[(\mathbf{Y} - \mathbf{XB})^\top (\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}^{-1}\right]\right\} \\ &= |\mathbf{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2} tr\left[\left(\mathbf{S} + (\mathbf{B} - \widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}})\right)\mathbf{\Sigma}^{-1}\right]\right\}, \end{aligned}$$

where $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^\top (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})$, $\widehat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$ (see Exercise 9).

The conjugate prior for this models is $\pi(\mathbf{B}, \mathbf{\Sigma}) = \pi(\mathbf{B}|\mathbf{\Sigma})\pi(\mathbf{\Sigma})$ where $\pi(\mathbf{B}|\mathbf{\Sigma}) \sim N_{K\times M}(\mathbf{B}_0, \mathbf{V}_0, \mathbf{\Sigma})$ and $\pi(\mathbf{\Sigma}) \sim IW(\mathbf{\Psi}_0, \alpha_0)$, that is,

$$\begin{aligned} \pi(\mathbf{B}, \mathbf{\Sigma}) \propto &|\mathbf{\Sigma}|^{-K/2} \exp\left\{-\frac{1}{2} tr\left[(\mathbf{B} - \mathbf{B}_0)^\top \mathbf{V}_0^{-1}(\mathbf{B} - \mathbf{B}_0)\mathbf{\Sigma}^{-1}\right]\right\} \\ &\times |\mathbf{\Sigma}|^{-(\alpha_0 + M + 1)/2} \exp\left\{-\frac{1}{2} tr\left[\mathbf{\Psi}_0 \mathbf{\Sigma}^{-1}\right]\right\}. \end{aligned}$$

The posterior distribution is given by

[5] vec denotes the vectorization operation, and $\otimes$ denotes the kronecker product.

[6] We can write down the former expression in a more familiar way using vectorization properties, $\underbrace{vec(Y)}_{\mathbf{y}} = \underbrace{(\mathbf{I}_M \otimes \mathbf{X})}_{\mathbf{Z}} \underbrace{vec(\mathbf{B})}_{\beta} + \underbrace{vec(\mathbf{U})}_{\mu}$, where $\mathbf{y} \sim N_{N\times M}(\mathbf{Z}\beta, \mathbf{\Sigma} \otimes \mathbf{I_N})$.

$$\begin{aligned}\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) &\propto p(\mathbf{Y}|\mathbf{B},\boldsymbol{\Sigma},\mathbf{X})\pi(\mathbf{B}|\boldsymbol{\Sigma})\pi(\boldsymbol{\Sigma})\\ &\propto |\boldsymbol{\Sigma}|^{-\frac{N+K+\alpha_0+M+1}{2}}\\ &\times \exp\left\{-\frac{1}{2}tr\left[(\boldsymbol{\Psi_0}+\mathbf{S}+(\mathbf{B}-\mathbf{B_0})^\top\mathbf{V_0^{-1}}(\mathbf{B}-\mathbf{B_0})\right.\right.\\ &\left.\left.+(\mathbf{B}-\widehat{\mathbf{B}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{B}-\widehat{\mathbf{B}}))\boldsymbol{\Sigma}^{-\mathbf{1}}\right]\right\}.\end{aligned}$$

Completing the squares on $\mathbf{B}$ and collecting the remaining terms in the bracket yields

$$\boldsymbol{\Psi}_0+\mathbf{S}+(\mathbf{B}-\mathbf{B}_0)^\top\mathbf{V}_0^{-1}(\mathbf{B}-\mathbf{B}_0)+(\mathbf{B}-\widehat{\mathbf{B}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{B}-\widehat{\mathbf{B}})=(\mathbf{B}-\mathbf{B}_n)^\top\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)+\boldsymbol{\Psi}_n,$$

where

$$\begin{aligned}\mathbf{B}_n =&(\mathbf{V}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{V}_0^{-1}\mathbf{B}_0+\mathbf{X}^\top\mathbf{Y})=(\mathbf{V}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{V}_0^{-1}\mathbf{B}_0+\mathbf{X}^\top\mathbf{X}\widehat{\mathbf{B}}),\\ \mathbf{V}_n =&(\mathbf{V}_0^{-1}+\mathbf{X}^\top\mathbf{X})^{-1},\\ \boldsymbol{\Psi}_n =&\boldsymbol{\Psi}_0+\mathbf{S}+\mathbf{B}_0^\top\mathbf{V}_0^{-1}\mathbf{B}_0+\widehat{\mathbf{B}}^\top\mathbf{X}^\top\mathbf{X}\widehat{\mathbf{B}}-\mathbf{B}_n^\top\mathbf{V}_n^{-1}\mathbf{B}_n.\end{aligned}$$

Thus, the posterior distribution can be written as

$$\begin{aligned}\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) \propto |\boldsymbol{\Sigma}|^{-K/2}\exp\left\{-\frac{1}{2}tr\left[(\mathbf{B}-\mathbf{B}_n)^\top\mathbf{V}_n^{-1}(\mathbf{B}-\mathbf{B}_n)\boldsymbol{\Sigma}^{-1}\right]\right\}\\ \times|\boldsymbol{\Sigma}|^{-\frac{N+\alpha_0+M+1}{2}}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Psi}_n\boldsymbol{\Sigma}^{-1}\right]\right\}.\end{aligned}$$

That is $\pi(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X}) = \pi(\mathbf{B}|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X})\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X})$ where $\pi(\mathbf{B}|\boldsymbol{\Sigma},\mathbf{Y},\mathbf{X}) \sim N_{K\times M}(\mathbf{B}_n,\mathbf{V}_n,\boldsymbol{\Sigma})$ and $\pi(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{X})\sim IW(\boldsymbol{\Psi}_n,\alpha_n)$ where $\alpha_n = N+\alpha_0$.

The marginal posterior for $\mathbf{B}$ is given by

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

The second line uses the inverse-Wishart distribution, the third line the Sylverter's theorem, and the last line is the kernel of a matrix t-distribution, that is, $\mathbf{B}|\mathbf{Y},\mathbf{X}\sim t_{K,M}(\mathbf{B}_n,\mathbf{V}_n,\boldsymbol{\Psi}_n)$ with α_n+1-M degrees of freedom.

Observe that $vec(\mathbf{B})$ has mean $vec(\mathbf{B}_n)$ and variance $(\mathbf{V}_n\otimes\boldsymbol{\Psi}_n)/(\alpha_n-M-1)$ based on its marginal distribution. On the other hand, the variance based on the conditional distribution is $\mathbf{V}_n\otimes\boldsymbol{\Sigma}$, where the mean of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Psi}_n/(\alpha_n-M-1)$.

The marginal likelihood is the following,

$$
\begin{aligned}
p(\mathbf{Y}) &= \int_{\mathcal{B}}\int_{\mathcal{S}} \left\{ (2\pi)^{-NM/2} |\mathbf{\Sigma}|^{-N/2} \exp\left\{ -\frac{1}{2} tr\left[\mathbf{S} + (\mathbf{B}-\widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B}-\widehat{\mathbf{B}}) \right] \mathbf{\Sigma}^{-1} \right\} \right. \\
&\quad \times (2\pi)^{-KM/2} |\mathbf{V}_0|^{-M/2} |\mathbf{\Sigma}|^{-K/2} \exp\left\{ -\frac{1}{2} tr\left[(\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1} (\mathbf{B}-\mathbf{B}_0) \mathbf{\Sigma}^{-1} \right] \right\} \\
&\quad \left. \times \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} |\mathbf{\Sigma}|^{-(\alpha_0+M+1)/2} \exp\left\{ -\frac{1}{2} tr\left[\mathbf{\Psi}_0 \mathbf{\Sigma}^{-1} \right] \right\} \right\} d\mathbf{\Sigma} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} \\
&\quad \times \int_{\mathcal{B}}\int_{\mathcal{S}} \left\{ |\mathbf{\Sigma}|^{-(\alpha_0+N+K+M+1)/2} \right. \\
&\quad \left. \exp\left\{ -\frac{1}{2} tr\left[\mathbf{S} + (\mathbf{B}-\widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B}-\widehat{\mathbf{B}}) + (\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1} (\mathbf{B}-\mathbf{B}_0) + \mathbf{\Psi}_0 \right] \mathbf{\Sigma}^{-1} \right\} \right\} d\mathbf{\Sigma} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathcal{B}} \left| \mathbf{S} + (\mathbf{B}-\widehat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B}-\widehat{\mathbf{B}}) + (\mathbf{B}-\mathbf{B}_0)^\top \mathbf{V}_0^{-1} (\mathbf{B}-\mathbf{B}_0) + \mathbf{\Psi}_0 \right|^{-(\alpha_n+K)/2} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathcal{B}} \left| (\mathbf{B}-\widehat{\mathbf{B}}_n)^\top \mathbf{V}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n) + \mathbf{\Psi}_n \right|^{-(\alpha_n+K)/2} d\mathbf{B} \\
&= (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathcal{B}} \left[|\mathbf{\Psi}_n| \times |\mathbf{I}_K + \mathbf{V}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n) \mathbf{\Psi}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n)^\top| \right]^{-(\alpha_n+K)/2} d\mathbf{B} \\
&= |\mathbf{\Psi}_n|^{-(\alpha_n+K)/2} (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \int_{\mathbf{B}} \left| \mathbf{I}_K + \mathbf{V}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n) \mathbf{\Psi}_n^{-1} (\mathbf{B}-\widehat{\mathbf{B}}_n)^\top \right|^{-(\alpha_n+1-M+K+M-1)/2} d\mathbf{B} \\
&= |\mathbf{\Psi}_n|^{-(\alpha_n+K)/2} (2\pi)^{-M(N+K)/2} |\mathbf{V}_0|^{-M/2} \frac{|\Psi_0|^{\alpha_0/2}}{2^{\alpha_0 M/2} \Gamma_M(\alpha_0/2)} 2^{M(\alpha_n+K)/2} \Gamma_M((\alpha_n+K)/2) \\
&\quad \times \pi^{MK/2} \frac{\Gamma_M((\alpha_n+1-M+M-1)/2)}{\Gamma_M((\alpha_n+1-M+K+M-1)/2)} |\mathbf{\Psi}_n|^{K/2} |\mathbf{V}_n|^{M/2} \\
&= \frac{|\mathbf{V}_n|^{M/2}}{|\mathbf{V}_0|^{M/2}} \frac{|\mathbf{\Psi}_0|^{\alpha_0/2}}{|\mathbf{\Psi}_n|^{\alpha_n/2}} \frac{\Gamma_M(\alpha_n/2)}{\Gamma_M(\alpha_0/2)} \pi^{-MN/2}.
\end{aligned}
$$

The third equality follows from having the kernel of a inverse-Wishart distribution, the fifth from the Silvester's theorem, and the seventh from having the kernel of a matrix t distribution.

Observe that this last expression is the multivariate case of the marginal likelihood of the univariate regression model. Taking into account that

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^{-1} &= \mathbf{A}^{-1}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1}\\ &= \mathbf{B}^{-1}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}\\ &= \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{B}^{-1},\end{aligned}$$

we can show that $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \mathbf{S} + (\hat{\mathbf{B}}-\mathbf{B}_0)^\top \mathbf{V}_n(\hat{\mathbf{B}}-\mathbf{B}_0)$ (see Exercise 7). Therefore, the marginal likelihood rewards fit (smaller sum of squares, $\mathbf{S}$), similarity between prior and sample information regarding location parameters, and information gains in variability from $\mathbf{V}_0$ to $\mathbf{V}_n$.

Given a matrix of regressors $\mathbf{X}_0$ for N_0 unobserved units, the predictive density of $\mathbf{Y}_0$ given $\mathbf{Y}$, $\pi(\mathbf{Y}_0|\mathbf{Y})$ is a matrix t distribution $T_{N_0,M}(\alpha_n - M + 1, \mathbf{X}_0\mathbf{B}_n, \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top, \boldsymbol{\Psi}_n)$ (see Exercise 6). Observe that the prediction is centered at $\mathbf{X}_0\mathbf{B}_n$, and the covariance matrix of $vec(\mathbf{Y}_0)$ is $\frac{(\mathbf{I}_{N_0}+\mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top)\otimes\boldsymbol{\Psi}_n}{\alpha_n - M - 1}$.

4.5 Computational examples

* ¿What is the probability that the Sun will rise tomorrow?

This is the most famaous Ricard Price's example developed in the Appendix of the Bayes' theorem paper [4]. 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 on a specified day rise, that is, a prior uniform distribution over (0,1), that is, a beta (1,1) distribution...

4.6 Summary: Chapter 4

4.7 Exercises: Chapter 4

1. Write in the canonical form the distribution of the Bernoulli example, and find the mean and variance of the sufficient statistic.
2. Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ 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.

3. Given a random sample $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ 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.
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.
5. Find the marginal likelihood in the normal/inverse-Wishart model.
6. Find the posterior predictive distribution in the normal/inverse-Wishart model, and show that $\mathbf{Y}_0|\mathbf{Y} \sim T_{N_0,M}(\alpha_n - M + 1, \mathbf{X}_0\mathbf{B}_n, \mathbf{I}_{N_0} + \mathbf{X}_0\mathbf{V}_n\mathbf{X}_0^\top, \mathbf{\Psi}_n)$.
7. 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.
8. 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 $\beta_{**} = \mathbf{X}_0\beta_n$.
9. Show that $(\mathbf{Y} - \mathbf{XB})^\top(\mathbf{Y} - \mathbf{XB}) = \mathbf{S} + (\mathbf{B} - \widehat{\mathbf{B}})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{B} - \widehat{\mathbf{B}})$ 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.

5

Simulation methods

Part II

Regression models: A GUIded tour

6

Univariate models

7

Multivariate models

8

Time series models

9

Panel data models

10

Bayesian model average

Part III

Recent developments: Theory, applications and programming

11

Hierarchical models

11.1 Direchlet processes

12

Causal inference

13

Machine learning

14

Spatial econometric models

15

Further topics

Bibliography

[1] M. Bayarri and J. Berger. P–values for composite null models. *Journal of American Statistical Association*, 95:1127–1142, 2000.

[2] M. J. Bayarri and J. Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

[3] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–416, 1763.

[4] 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.

[5] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10, 2018.

[6] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, third edition edition, 1993.

[7] J. Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[8] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[9] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.

[10] Peter J Bickel and Joseph A Yahav. Some contributions to the asymptotic theory of bayes solutions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11(4):257–276, 1969.

[11] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 1976.

[12] George EP Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

[13] V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346, 2003.

[14] Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

[15] Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[16] M. Clyde and E. George. Model uncertatinty. *Statistical Science*, 19(1):81–94, 2004.

[17] A. P. Dawid, M. Musio, and S. E. Fienberg. From statistical evidence to evidence of causality. *Bayesian Analysis*, 11(3):725–752, 2016.

[18] de Finetti. Foresight: its logical laws, its subjective sources. In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*. Krieger, New York, 1937. p.55–118.

[19] M. H. DeGroot. *Probability and statistics*. Addison-Wesley Publishing Co., London, 1975.

[20] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[21] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[22] R. Fisher. *Statistical Methods for Research Workers*. Hafner, New York, 13th edition, 1958.

[23] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[24] Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

[25] A. Gelman and X. Meng. Model checking and model improvement. In Gilks, Richardson, and Speigelhalter, editors, *In Markov chain Monte Carlo in practice*. Springer US, 1996. Chapter 6, pp. 157–196.

[26] A. Gelman, X. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[27] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[28] Andrew Gelman and Guido Imbens. Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research, 2013.

[29] S Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[30] I. J. Good. The bayes/non bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419):597–606, September 1992.

[31] S. N. Goodman. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine*, 130(12):995–1004, 1999.

[32] W. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109, 1970.

[33] H. Jeffreys. Some test of significance, treated by the theory of probability. *Proccedings of the Cambridge philosophy society*, 31:203–222, 1935.

[34] H. Jeffreys. *Theory of Probability*. Oxford University Press, London, 1961.

[35] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[36] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[37] R. Kass. Statistical inference: the big picture. *Statistical science*, 26(1):1–9, 2011.

[38] Robert E. Kass and Adrian E. Raftery. Bayes factorss. *Journal of American Statistical Association*, 90(430):773–795, 1995.

[39] Gary M Koop. *Bayesian econometrics*. John Wiley & Sons Inc., 2003.

[40] P. Laplace. *Théorie Analytique des Probabilités*. Courcier, 1812.

[41] Pierre Simon Laplace. Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*, 1774.

[42] E.L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, second edition edition, 2003.

[43] D. V. Lindley. The philosophy of statistics. *The Statistician*, 49(3):293–337, 2000.

[44] D. V. Lindley and L. D. Phillips. Inference for a Bernoulli process (a Bayesian view). *American Statistician*, 30:112–119, 1976.

[45] Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[46] Sharon Bertsch McGrayne. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C.* Yale University Press, 2011.

[47] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys*, 21:1087–1092, 1953.

[48] J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A*, 231:289–337, 1933.

[49] G. Parmigiani and L. Inoue. *Decision theory principles and approaches.* John Wiley & Sons, 2008.

[50] Luis Pericchi and Carlos Pereira. Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics*, 2015.

[51] Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli. Dynamic linear models. In *Dynamic Linear Models with R*, pages 31–84. Springer, 2009.

[52] A. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

[53] F. Ramsey. Truth and probability. In Routledge and Kegan Paul, editors, *The Foundations of Mathematics and other Logical Essays.* New York: Harcourt, Brace and Company, London, 1926. Ch. VII, p.156–198.

[54] L. J. Savage. *The foundations of statistics.* John Wiley & Sons, Inc., New York, 1954.

[55] Robert Schlaifer and Howard Raiffa. *Applied statistical decision theory.* Wiley New York, 1961.

[56] Thomas Sellke, MJ Bayarri, and James O Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[57] Steve Selvin. A problem in probability (letter to the editor). *The American Statistician*, 11(1):67–71, 1975.

[58] Steve Selvin. A problem in probability (letter to the editor). *The American Statistician*, 11(3):131–134, 1975.

[59] A. F. M. Smith. A General Bayesian Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological).*, 35(1):67–75, 1973.

[60] Stephen Stigler. Richard price, the first bayesian. *Statistical Science*, 33(1):117–125, 2018.

[61] Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728, 1994.

[62] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[63] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p–values: context, process and purpose. *The American Statistician*, 2016.

[64] A. Zellner. *Introduction to Bayesian inference in econometrics*. John Wiley & Sons Inc., 1996.

[65] S. Ziliak. Guinnessometrics; the Economic Foundation of student's t. *Journal of Economic Perspectives*, 22(4):199–216, 2008.