Bayesian Statistics: Lecture Notes

Alessandra Guglielmi
Department of Mathematics
Politecnico di Milano

alessandra.guglielmi@polimi.it

October 2024

The author of these notes is Alessandra Guglielmi, instructor of the course *Bayesian Statistics* for the M.Sc. program in Mathematical Engineering at Politecnico di Milano. The notes are based on material from the following textbooks:

- Rosner, G. L., Laud, P. W., and Johnson, W. O. (2021). *Bayesian thinking in Biostatistics*. CRC Press.
- Jackman, S. (2009). *Bayesian analysis for the Social Sciences*. John Wiley & Sons.

Some of the examples reported here are from
Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, Springer.

You must give appropriate credit if you change these notes before using them.

Acknowledgements
I am grateful to Michele Russo, who took the notes when attending the course, and to Alessandro Carminati who helped me to write them in Latex.

Contents

1 Basics of Bayesian Inference

We often use probabilities informally to express our information and beliefs about unknown quantities. However, the use of probabilities to express information can be made formal: in a precise mathematical sense, it can be shown that probabilities can numerically represent a set of rational beliefs or *previous* information, and that the Bayes rule provides a rational method for updating beliefs in light of new information. The process of inductive learning via the Bayes rule is referred to as **Bayesian inference**. Bayesian statistics is based on the premise that all uncertainty should be modeled using probabilities and that statistical inferences should be logical conclusions based on the law of probability. Bayesian statistics is model-based, i.e., it is based on probability models for data. These models involve parameters that are presumed to be related to characteristics of the sampled populations. However, parameters can never be known with absolute certainty (unless we sample the whole population); they may not have physical interpretations; models are useful approximations to some truth.
In addition to their formal interpretation as a tool for induction, Bayesian inference provides:

- Parameter estimates with good statistical properties.
- Parsimonious descriptions of observed data.
- Predictions for missing data and forecasts of future data.
- A computational framework for model estimation, selection and validation.

Methods of Bayesian statistics emphasize the model-based approach to data analysis, i.e., we need to specify a probabilistic mechanism that might have generated the data.

1.1 Bayesian Learning: Likelihood, Prior and Posterior

Statistical induction is the process of learning about the general characteristics of a population from a subset of members of that population. Numerical values of population characteristics are typically expressed in terms of a parameter θ, while $\boldsymbol{y}$ is the numerical description of the sample. Before a dataset is observed, the numerical values of both the population characteristics and the dataset are uncertain. After a dataset $\boldsymbol{y}$ is obtained, the information it contains can be used to decrease our uncertainty about the population characteristics. This change is quantified by the conditional distribution of θ, given $\boldsymbol{y}$.
The **sample space** $\mathcal{Y}$ is the set of all possible datasets, from which a single dataset $\boldsymbol{y}$ will result. The **parameter space** Θ is the set of possible parameter values, from which we hope to identify the value that best represents the true population characteristics. Bayesian learning is assigned by assigning the joint beliefs about $\boldsymbol{Y}$ and θ, expressed in terms of probability distributions over $\mathcal{Y}$ and Θ. Note that $\boldsymbol{Y}$ is the random vector representing data before recording the observed dataset $\boldsymbol{y}$. We use instead the same symbol θ to denote the random variable and the value recorded.
Under the Bayesian approach, unlike the classical approach, there are two random elements, $(\boldsymbol{Y}, \theta)$, and we typically assign their joint distribution as follows:

1. For each numerical value $\theta \in \Theta$, our **prior distribution** $\pi(\theta)$ describes our belief that θ represents the true population characteristics.
2. For each $\theta \in \Theta$ and $\boldsymbol{y} \in \mathcal{Y}$, our **likelihood** $f(\boldsymbol{y}|\theta)$ describes our belief that $\boldsymbol{y}$ would be the outcome of our study if we knew θ to be true.

Once we obtain the data $\boldsymbol{y}$, the last step is to update our beliefs about θ:

3. For each numerical value of $\theta \in \Theta$, our **posterior distribution** $\pi(\theta|\boldsymbol{y})$ describes our belief that θ is the true value, having observed $\boldsymbol{y}$.

The posterior distribution is obtained from the prior distribution and the likelihood via the Bayes rule:

$$\pi(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)\,\pi(\theta)}{\int_{\Theta} p(\boldsymbol{y}|\theta)\,\pi(\theta)\,\mathrm{d}\theta}$$

The Bayesian approach is such that the prior belief is updated via observed data and yields posterior distribution. Moreover, it suggests that scientific inference is based on two parts: one depends on the scientist's subjective opinion and understanding of the phenomenon under study **before an experiment** was performed, and the other depends on the observed data the scientist has obtained from the experiment under study itself.
Given a statistical model for the data, the Bayesian approach mandates an additional probability model for all the parameters in the data model. We model the uncertainty about parameters typically by using scientific expert information (*knowledge based info*), but this information must be obtained independently of the data being analyzed (though this comment will be revised). We need to assume independent information to set the prior distribution of the parameter θ, i.e., from existing literature on the data similar to the current data being analyzed.
Bayesian inference is totally based on the posterior distribution $\pi(\theta|\boldsymbol{y})$, computing summaries of this distribution:

- posterior mean $E[\theta|y_1,\ldots,y_n]$
- posterior variance $\text{Var}[\theta|y_1,\ldots,y_n]$
- interval estimate $C : \mathbb{P}(\theta \in C|y_1,\ldots,y_n) \geq 0.95$

However, in all real-world applications, we will need computational methods to derive these estimates, or, more generally, to simulate a sample from the posterior distribution. Hence, we will introduce later in this course Markov chain Monte Carlo (MCMC) methods. The key idea is that if we are not able to derive an iid sample from the *posterior distribution*, we aim at building a Markov chain with limiting distribution given by the posterior distribution itself, and then we apply the Ergodic Theorem to approximate integrals from the posterior.

1.2 Bayes' Theorem for Dominated Models

1.2.1 Bayes' Theorem for Events

We review Bayes' theorem for an event-space partition.

Theorem 1.1. *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and $H_1,\ldots,H_K$ be a (measurable) partition of Ω such that $\mathbb{P}(H_k) > 0$ for all $k = 1,\ldots,K$. Moreover, consider an event $E \in \mathcal{F}$ such that $\mathbb{P}(E) > 0$. We have*

$$\mathbb{P}(H_j|E) = \frac{\mathbb{P}(E|H_j)\,\mathbb{P}(H_j)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|H_j)\,\mathbb{P}(H_j)}{\sum_{k=1}^{K}\mathbb{P}(E|H_i)\,\mathbb{P}(H_k)} \quad \text{for any } j = 1,\ldots,K.$$

The theorem can be generalized to a denumerable partition of Ω, i.e., $K = \infty$. The proof follows by the definition of the conditional probability.

1.2.2 Bayes' Theorem for Random Variables

We now introduce the Bayes' theorem for two random variables.

Theorem 1.2. *Let Y and θ be two random variables such that $Y|\theta \sim f(\cdot|\theta)$ and $\theta \sim \pi(\cdot)$ where f and π are densities. Then the posterior density of θ, that is the conditional density of θ, given $Y = y$ is computed as*

$$\pi(\theta|y) = \frac{f(y|\theta)\,\pi(\theta)}{\int_\Theta f(y|\theta)\,\pi(\theta)\,d\theta}$$

where Θ is the space of possible values of θ.

The equality follows from the definition of a conditional density, and the denominator is simply the marginal density for Y, calculated as the integral of the joint density of Y,θ, $f(y|\theta)\,\pi(\theta)$.

1.2.3 Bayes' Theorem for Dominated Models

We can now finally introduce the Bayesian paradigm in the usual form. Consider a finite (so far) sequence of random variables $\boldsymbol{Y}_n = Y_1, \ldots, Y_n$ over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let P_θ be the distribution of $\boldsymbol{Y}_n|\theta$ and $\theta \in \Theta \subset \mathbb{R}^p$. In the **Bayesian approach**, parameter θ, representing unknown characteristics of distribution of data $\boldsymbol{Y}_n$ is a random vector distributed according to a probability measure π on $(\Theta, \mathcal{B}(\theta))$ which is called the **prior distribution**. We assume here that P_θ, the conditional distribution of data $\boldsymbol{Y}_n$, given θ, also called the **likelihood, as a function of** θ, has density $f(\cdot|\theta)$. We say that we have a **dominated model**. The **posterior distribution** is the conditional law of θ given the realization $\boldsymbol{y}_n$ of $\boldsymbol{Y}_n$. The posterior distribution can be derived through Bayes' theorem.

Theorem 1.3. *We have*

$$\mathbb{P}(\theta \in B|\boldsymbol{Y} = \boldsymbol{y}) = \frac{\int_B f(\boldsymbol{y}|\theta)\,\pi(d\theta)}{\int_\Theta f(\boldsymbol{y}|\theta)\,\pi(d\theta)} \text{ for all } B \in \mathcal{B}. \tag{1}$$

If π has a density, denoted by $\pi(\theta)$, then the posterior can be computed as

$$\pi(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)\,\pi(\theta)}{\int_\Theta f(\boldsymbol{y}|\theta)\,\pi(\theta)\,d\theta} \tag{2}$$

The denominator of these formulas above is the marginal density of $\boldsymbol{Y}_n$, i.e.,

$$m_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\theta)\,\pi(\theta)\,d\theta \tag{3}$$

The set of values $\boldsymbol{y}$ where density $m_{\boldsymbol{Y}}(\boldsymbol{y})$ is equal to zero has zero-measure with respect to the marginal distribution of $\boldsymbol{Y}_n$.

Remark. *The most simple framework for model-based inference is to assume an infinite sequence of random variables $Y_1, Y_2, Y_3, \ldots$, representing an infinite (or very large) number of observations, and assume that, for any n*

$$Y_1, \ldots, Y_n|\theta \overset{iid}{\sim} f_\theta$$

where f_θ is the conditional density of any observation Y_i. In this case, the posterior density in (2) *assumes the following form*

$$\pi(\theta|\boldsymbol{y}) = \frac{\prod_{i=1}^n f_\theta(y_i)\pi(\theta)}{\int_\Theta \prod_{i=1}^n f_\theta(y_i)\pi(\theta)\,d\theta}.$$

Observe that the posterior is proportional to the product of the conditional distribution of data $Y_1, \ldots, Y_n$ given θ, i.e., the likelihood, $\prod_{i=1}^n f_\theta(y_i)$ and the prior $\pi(\theta)$.

If we have two datasets, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, available at different times. The Bayesian learning may be carried on in two alternative ways:

- We could choose a prior for θ, update such prior through Bayes's theorem using only $\boldsymbol{y}_1$; then consider the posterior of θ given $\boldsymbol{y}_1$ as a new prior for θ to be updated using only $\boldsymbol{y}_1$ (assuming the same likelihood in both cases).
- On the other hand, we could choose to update the prior for θ using both $\boldsymbol{y}_1, \boldsymbol{y}_2$ at once.

However, it is easy to show that (if we start with the same prior) these two methods give the same final posterior, if data $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are conditionally independent, given θ. Indeed, the posterior corresponding to the second method can be written as

$$\pi(\theta|\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \frac{\mathcal{L}(\boldsymbol{Y}_1, \boldsymbol{Y}_2|\theta)\,\pi(\theta)}{\mathcal{L}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)} = \frac{\mathcal{L}(\boldsymbol{Y}_1|\theta)\,\mathcal{L}(\boldsymbol{Y}_2|\theta)\,\pi(\theta)}{\mathcal{L}(\boldsymbol{Y}_1)\,\mathcal{L}(\boldsymbol{Y}_2|\boldsymbol{Y}_1)} = \frac{\mathcal{L}(\boldsymbol{Y}_1|\theta)\,\pi(\theta)}{\mathcal{L}(\boldsymbol{Y}_1)} \times \frac{\mathcal{L}(\boldsymbol{Y}_2|\theta)}{\mathcal{L}(\boldsymbol{Y}_2|\boldsymbol{Y}_1)} = \frac{\pi(\theta|\boldsymbol{Y}_1)\,\mathcal{L}(\boldsymbol{Y}_2|\theta)}{\mathcal{L}(\boldsymbol{Y}_2|\boldsymbol{Y}_1)}.$$

The right handside describes the distribution from the first method.

1.3 Main Inferential Problems

1.3.1 Point and Interval Estimation

We have seen how to compute the posterior density once we have assigned the conditional distribution of the data given parameter (likelihood) and the prior distribution of the parameter itself. We now wonder how to summarize the posterior distribution we get via Bayes' theorem: in other words, we want to understand what type of estimates we can consider for the parameter θ. We can provide point estimates such as the **posterior mean** or the **posterior median** of all component of the parameter vector θ.
We can also consider interval estimates, which in the Bayesian framework are denoted by **credible intervals/regions**. A credible interval is an interval $C \subset \Theta \subset \mathbb{R}$ where the scalar parameter θ belongs with a large posterior probability $\mathbb{P}_\pi(\theta \in C|\boldsymbol{y}) \geq 1-\alpha$ with $\alpha \in (0,1)$, α small. This definition is consistent with the existence of many different credible intervals of the same credibility level $1-\alpha$.
We define a **credible region** if θ is multidimensional, $C \subset \Theta$ and $\mathbb{P}_\pi(\theta \in C|\boldsymbol{y}) \geq 1-\alpha$.
When θ is unidimensional and typically absolutely continuous, methods for defining a suitable credible interval include:

- Choosing the **equal-tailed credible interval** $C = \left(q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}\right)$ where $1-\alpha$ is the level of the interval and q_α denotes the α-quantile of the posterior density.
- Choosing the narrowest interval, which for a unimodal distribution will involve choosing those values of highest probability density (**highest probability density interval**): $C = \{\theta|\pi(\theta|\boldsymbol{y}) \geq k\}$ where k is such that $\mathbb{P}_\pi(\theta \in C|\boldsymbol{y}) \geq 1-\alpha$ and $1-\alpha$ is the level of the interval.

1.3.2 Hypothesis Testing

The goal of this section is to test the hypotheses $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\{\Theta_0, \Theta_1\}$ is a disjoint partition of Θ and we assume that $\dim \Theta_0 = \dim \Theta_1$ for the sake of simplicity (see the TA class for $H_0 : \theta = \theta_0$).
We denote by $\pi_0 := \pi(\Theta_0)$, $\pi_1 := \pi(\Theta_1) = 1-\pi_0$ the prior mass of the alternative hypotheses. Let g_0 be the prior density of θ if H_0 is true and g_1 the prior density of θ if H_1 is true. For instance, we could assign one single density g to the whole parametric space Θ and denote by g_i the density g restricted to θ_i, for $i = 0, 1$. Hence, the expression for the prior density π for θ is

$$\pi(\theta) = \pi_0 g_0(\theta)\, \mathbb{1}_{\Theta_0}(\theta) + (1-\pi_0)\, g_1(\theta)\, \mathbb{1}_{\Theta_1}(\theta).$$

Here $\mathbb{1}_A(x) = 1$ if $x \in A$ and $= 0$ otherwise.
Let $f(\boldsymbol{y}|\theta)$ be the likelihood for data $\boldsymbol{y}$. By (2), the posterior density is then

$$\pi(\theta|\boldsymbol{y}) = \begin{cases} \frac{\pi_0 f(\boldsymbol{y}|\theta) g_0(\theta)}{m(\boldsymbol{y})} & \text{if } \theta \in \Theta_0 \\ \frac{(1-\pi_0) f(\boldsymbol{y}|\theta) g_1(\theta)}{m(\boldsymbol{y})} & \text{if } \theta \in \Theta_1 \end{cases}$$

where $m(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\theta)\, \pi(\theta)\, \mathrm{d}\theta = \pi_0 \int_{\Theta_0} f(\boldsymbol{y}|\theta)\, g_0(\theta)\, \mathrm{d}\theta + (1-\pi_0) \int_{\Theta_1} f(\boldsymbol{y}|\theta)\, g_1(\theta)\, \mathrm{d}\theta$ is the marginal density of the data. It is now possible to compute the posterior probability of Θ_0:

$$\mathbb{P}_\pi(\theta \in \Theta_0|\boldsymbol{y}) = \pi_0 \frac{\int_{\Theta_0} f(\boldsymbol{y}|\theta)\, g_0(\theta)\, \mathrm{d}\theta}{m(\boldsymbol{y})}.$$

The posterior probability of Θ_1 can be computed as

$$1 - \mathbb{P}_\pi(\theta \in \Theta_0|\boldsymbol{y}) = \pi_1 \frac{\int_{\Theta_1} f(\boldsymbol{y}|\theta)\, g_1(\theta)\, d\theta}{m(\boldsymbol{y})}.$$

One first criteria is to choose H_0 if $\mathbb{P}_\pi(\theta \in \Theta_0|\boldsymbol{y}) \geq 0.5$, or H_1 otherwise. If we were very concerned about making a mistake of, say, select H_0 when in fact H_1 is true, we might want to make it more

difficult to select H_0 by increasing the threshold for model selection from 0.5 to a larger value, perhaps 0.95. In this case, we would be at least 95% certain that H_0 is true. We could select H_0 if $\mathbb{P}_\pi(\theta \in \Theta_0|\boldsymbol{y}) \geq k$ for some $k > 0.5$. Making a decision based on this criterion is equivalent to selecting H_0 if we have that $\frac{\mathbb{P}_\pi(\theta\in\Theta_0|\boldsymbol{y})}{\mathbb{P}_\pi(\theta\in\Theta_1|\boldsymbol{y})} > \frac{k}{1-k}$, where the quotient $\frac{\mathbb{P}_\pi(\theta\in\Theta_0|\boldsymbol{y})}{\mathbb{P}_\pi(\theta\in\Theta_1|\boldsymbol{y})}$ is called **posterior odds**. The posterior odds can also be written as follows:

$$\frac{\mathbb{P}_\pi(\theta \in \Theta_0|\boldsymbol{y})}{\mathbb{P}_\pi(\theta \in \Theta_1|\boldsymbol{y})} = \frac{\pi_0 \frac{\int_{\Theta_0} f(\boldsymbol{y}|\theta) g_0(\theta) \mathrm{d}\theta}{m(\boldsymbol{y})}}{(1-\pi_0)\frac{\int_{\Theta_1} f(\boldsymbol{y}|\theta) g_1(\theta) \mathrm{d}\theta}{m(\boldsymbol{y})}} = \frac{\pi_0}{1-\pi_0} \times \frac{\int_{\Theta_0} f(\boldsymbol{y}|\theta)\, g_0(\theta)\, \mathrm{d}\theta}{\int_{\Theta_1} f(\boldsymbol{y}|\theta)\, g_1(\theta)\, \mathrm{d}\theta} = \frac{\mathbb{P}_\pi(\theta \in \Theta_0)}{\mathbb{P}_\pi(\theta \in \Theta_1)} \times BF_{01}$$

Note that the first factor in the right handside is the prior odds, while the second is referred to as the **Bayes factor**, denoted by BF_{01}, which is the factor by which the prior odds can be multiplied to be updated to the posterior odds. Observe that the Bayes factor can also be seen as the ratio of marginal densities of the data under H_0 and H_1:

$$BF_{01} = \frac{\int_{\Theta_0} f(\boldsymbol{y}|\theta)\, g_0(\theta)\, \mathrm{d}\theta}{\int_{\Theta_1} f(\boldsymbol{y}|\theta)\, g_1(\theta)\, \mathrm{d}\theta}.$$

The Bayes factor was proposed by Harold Jeffreys as a way to run hypothesis tests in the Bayesian framework. Jeffreys also suggested some levels of evidence to aid interpretation of the Bayes factor, which are summarized in the following table.

$1 < BF_{01} < 10^{\frac{1}{2}}$	There is evidence in favor of H_0, but it is not worth more than a bare mention.
$10^{\frac{1}{2}} < BF_{01} < 10$	There is substantial evidence in favor of H_0
$10 < BF_{01} < 10^2$	There is strong evidence in favor of H_0.
$BF_{01} > 10^2$	Evidence in favor of H_0 is decisive.

Remark. *Observe that the Bayes factor of H_1 versus H_0 is such that $BF_{10} = \frac{1}{BF_{01}}$.*

1.4 Posterior Predictive Distributions

Inference for unknowns in a model given data straightforwardly leads to prediction of quantities that involve future observations. Let $Y_1, \ldots, Y_n, \ldots$ be the (infinite) sequence of random variables representing observations. By (**posterior**) **predictive distribution** we mean the law $\mathcal{L}(Y_{n+1}, \ldots, Y_{n+m}|Y_1, \ldots, Y_n)$ for some m, given data $Y_1, \ldots, Y_n$. In particular, the one-step ahead posterior predictive distribution is the conditional law of Y_{n+1} given $Y_1, \ldots, Y_n$ and it represents the Bayesian forecast of Y_{n+1} on the basis of $Y_1, \ldots, Y_n$. If we now assume that $Y_1, \ldots, Y_n, Y_{n+1}|\theta \overset{\text{iid}}{\sim} f_\theta$, f_θ being the conditional density, we can compute the **posterior predictive density** as

$$m_{Y_{n+1}|Y_1,\ldots,Y_n}(y|y_1,\ldots,y_n) = \frac{m_{Y_1,\ldots,Y_{n+1}}(y_1,\ldots,y_n,y)}{m_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n)} = \frac{\int_\Theta f(y_1,\ldots,y_n,y|\theta)\,\pi(\theta)\,\mathrm{d}\theta}{\int_\Theta f(y_1,\ldots,y_n|\theta)\,\pi(\theta)\,\mathrm{d}\theta} \tag{4}$$

$$= \frac{\int_\Theta \prod_{i=1}^n f(y_i|\theta) f(y|\theta)\,\pi(\theta)\,\mathrm{d}\theta}{\int_\Theta \prod_{i=1}^n f(y_i|\theta)\pi(\theta)\,d\theta} = \int_\Theta f(y|\theta)\,\pi(\theta|y_1,\ldots,y_n)\,\mathrm{d}\theta \tag{5}$$

The first equality in (4) is the definition of the conditional density of Y_{n+1}, given $Y_1, \ldots, Y_n$, while the first equality in (5) follows because of conditional independence of all data. Hence, in the (conditional) iid context, the posterior predictive marginal density is the integral of the conditional distribution of one observation, with respect to the posterior of θ.

Example 1.1. *Consider data regarding the blood pressure of war veterans collected on a sample of size $n = 404$. Let us denote by Y_i the binary random variable indicating whether the i-th veteran has uncontrolled hypertension ($y_i = 1$) or not ($y_i = 0$). Moreover, let us assume the following model:*

$$\begin{aligned} Y_1, \dots, Y_{404} | \theta &\overset{iid}{\sim} Be(\theta) \\ \theta &\sim \mathcal{U}([0,1]). \end{aligned}$$

Suppose now that we are interested in computing the probability that a new war veteran has uncontrolled hypertension given that $\sum_{i=1}^{404} y_i = 184$. It is easy to show that

$$\pi(\theta|\boldsymbol{y}) = Beta(185, 221).$$

Such quantity can be computed through the posterior predictive distribution as follows:

$$\mathcal{L}(Y_{n+1}|Y_1 = y_1, \dots, Y_n = y_n) = \int_\Theta \mathcal{L}(Y_{n+1}|\theta)\, \pi(\theta|\boldsymbol{y})\, d\theta.$$

In particular,

$$\mathbb{P}(Y_{n+1} = 1|Y_1 = y_1, \dots, Y_n = y_n) = \int_\Theta \mathbb{P}(Y_{n+1} = 1|\theta)\, \pi(\theta|\boldsymbol{y})\, d\theta = \int_\Theta \theta \pi(\theta|\boldsymbol{y})\, d\theta = \mathbb{E}[\theta|\boldsymbol{y}] \approx 0.4557.$$

1.5 Exchangeability

We now introduce the notion of exchangeability for a sequence of random variables, which will lead to a reinterpretation of the Bayesian paradigm thanks to the statement of de Finetti's representation theorem.

Definition 1.1 (Exchangeability)**.** *The random vector $(Y_1, \dots, Y_n)$ is **exchangeable** if*

$$\mathcal{L}(Y_1, \dots, Y_n) = \mathcal{L}\left(Y_{\pi(1)}, \dots, Y_{\pi(n)}\right)$$

*for any permutation π of $(1, \dots, n)$. Moreover, the sequence of random variables $(Y_n)_{n\geq 1}$ is **exchangeable** if $(Y_1, \dots, Y_n)$ is exchangeable for any $n \geq 1$.*

Remark. *If $(Y_1, \dots, Y_n)$ is exchangeable we have that:*

$$\begin{aligned} \mathcal{L}(Y_1) &= \mathcal{L}(Y_i) \ \text{ for all } i \\ \mathcal{L}(Y_1, Y_2) &= \mathcal{L}(Y_i, Y_j) \ \text{ for all } i, j \geq 1, i \neq j \\ \mathcal{L}(Y_1, Y_2, Y_3) &= \mathcal{L}(Y_i, Y_j, Y_k) \ \text{ for all } i, j, k \geq 1, i, j, k \text{ different} \\ &\vdots \end{aligned}$$

Such condition then implies that the order with which data are recorded is irrelevant for inferential purposes. Therefore, exchangeability is a weak assumption that translates lack of information to a condition of symmetry.

Theorem 1.4 (de Finetti's representation)**.** *Let $(Y_n)_{n\geq 1}$ be a sequence of binary random variables. Such sequence is exchangeable if and only if there exists a probability measure F on $([0,1], \mathcal{B}([0,1]))$ such that*

$$\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n] = \int_{[0,1]} \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n - \sum_{i=1}^n y_i} F(d\theta) \text{ for all } (y_1, \dots, y_n) \in \{0,1\}^n \text{ and all } n.$$

In a nutshell, (one direction of) the de Finetti's theorem states that, if $(Y_n)_{n\geq 1}$ is exchangeable, then there exists a random variable θ such that $Y_1, \dots, Y_n | \theta \overset{\text{iid}}{\sim} \text{Be}(\theta)$ for all $n \geq 1$ with $\theta \sim F$. Therefore, there is a formal equivalence between exchangeable trials of the same phenomenon and trials designated as independent and with a fixed and unknown probability (which are those of the Bayesian framework).

Remark. *This result can be extended to more general sequences of random variables.*

See Regazzini (1996) and Schervish (2012).

1.6 Specifying Prior Distributions

The prior distribution should quantify external knowledge containing information outside the data being modeled and analyzed. The specification of priors is extremely important in Bayesian statistics and it deserves special attention, the main reason being the great potential for misspecification of priors, which can easily lead to what has been termed *garbage in, garbage out*. There are basically two kinds of prior:

- If there is scientifically relevant information available (there is always info), then it is important to incorporate it into the prior and in this case we have an **informative prior**.
- If such information is not available, the prior is generally specified so as to have a minimal impact on the Bayesian analysis and in this case we have a **reference prior** (or **convenience prior**).

1.6.1 Reference Priors

By reference prior we mean any prior that is not chosen for the information that it models. Rather, it is chosen to provide a common base for people to evaluate data.

Remark (Non-informative priors)**.** *Historically, there has been considerable effort spent on the development of so-called non-informative priors. The name stems from the fact that these priors are meant to have little influence on the posterior distribution. However, such terminology is strongly misleading: although priors that have little effect on the posterior do exist, there is really no such thing as a non-informative prior. Indeed, all priors express information about the parameters and often the information expressed by these priors is uniquely uninspired in the sense that nobody would use them to make decisions in the absence of data. Some may automatically believe from the name that a non-informative prior is genuinely not informative, since no scientific input has been applied. While there are situations where priors that are so named would have little effect on the posterior, there are also examples of such priors that are in fact, disinformative, meaning that they convey a form of disinformation that is contrary to the scientific context.*

One sort of reference prior is the **flat prior**, which is given by $\pi(\theta) = c$ for all $\theta \in \Theta$ where c is a positive constant; the prior is flat over the entire parameter space. Observe that if Θ is not bounded then the prior is an improper distribution as we have $\int_\Theta \pi(\theta)\,\mathrm{d}\theta = +\infty$ regardless of the choice for the constant c. Moreover, given a likelihood $f(y|\theta)$ if we choose a flat prior $\pi(\theta) = c$ then the marginal density for the data $m(y) = \int_\Theta f(y|\theta)\,\pi(\theta)\,\mathrm{d}\theta$ may not exist. However, the virtue of a flat prior is that it is often easily overwhelmed by the data. For instance, using $\pi(\theta) = c$ we have that Bayes theorem gives:

$$\pi(\theta|y) = \frac{f(y|\theta)\,\pi(\theta)}{\int_\Theta f(y|\theta)\,\pi(\theta)\,\mathrm{d}\theta} = \frac{f(y|\theta)}{\int_\Theta f(y|\theta)\,\mathrm{d}\theta}$$

Therefore, the posterior is simply a normalization of the likelihood as a function of θ.

Remark. *This is not true in general: indeed, not all likelihoods can be normalized as they may not have finite integrals with respect to θ.*

Example 1.2 (Ex. 6.1, Rosner et al.)**.** *Assume independent normally distributed data with unknown mean θ and known precision τ_0: $Y_1, \ldots, Y_n|\theta \overset{iid}{\sim} \mathcal{N}\left(\theta, \frac{1}{\tau_0}\right)$. Then the likelihood is given by $f(|\theta) \propto \mathrm{e}^{-\frac{\tau_0}{2}\sum_{i=1}^n (y_i-\theta)^2} \propto \mathrm{e}^{-\frac{n\tau_0}{2}(\overline{y}-\theta)^2}$. If we assume a flat prior we have that the posterior is given by:*

$$\pi(\theta|y_1, \ldots, y_n) = \frac{\mathrm{e}^{-\frac{n\tau_0}{2}(\overline{y}-\theta)^2} \times c}{\int_\Theta \mathrm{e}^{-\frac{n\tau_0}{2}(\overline{y}-\theta)^2} \times c\; d\theta} = \frac{1}{\sqrt{2\pi\frac{1}{n\tau_0}}}\mathrm{e}^{-\frac{n\tau_0}{2}(\overline{y}-\theta)^2}$$

so that $\theta|y_1, \ldots, y_n \sim \mathcal{N}\left(\overline{y}, \frac{1}{n\tau_0}\right)$. Therefore we have that $\mathbb{E}[\theta|y_1, \ldots, y_n] = \overline{y}$ (which coincides with the maximum likelihood estimate of θ); in this case, the posterior 95% credibility interval (centered at

$\bar{y}$) is given by $I = \left[\overline{y} \pm 1.96 \frac{\sigma_0}{\sqrt{n}}\right]$ *where* $\sigma_0 = \frac{1}{\sqrt{\tau_0}}$ *(which is the standard formula for the confidence interval for the mean of a Gaussian population with known variance).*

In the previous example we have obtained the standard frequentist formulas for point and interval estimation by choosing a flat prior. Indeed, one of the definitions of a reference prior is that it leads to Bayesian (posterior) point and interval estimates, as well as predictive inferences, that are the same as a frequentist would obtain without a prior. In this sense, the selected prior is often deemed not to affect the posterior "adversely".

Example 1.3 (Ex. 6.3, Rosner et al.)**.** *Let* $\theta \in (0,1)$ *be the success probability in a sequence of Bernoulli trials, i.e. the probability of an individual of being infected. Consider an improper prior on the unit interval:* $\pi(\theta) = \theta^{-1}(1-\theta)^{-1}\mathbb{I}_{(0,1)}(\theta)$*. This can be seen as the beta*(ϵ, ϵ) *distribution with* ϵ *small. We change the parameterization, considering, as the new parameter*

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

Then, the density of γ *is given by:*

$$\pi_\gamma(\gamma) = \pi_\theta\left(g^{-1}(\gamma)\right)\left|\frac{d}{d\gamma} g^{-1}(\gamma)\right| = 1 \quad \gamma \in \mathbb{R}.$$

Thus, our initial prior corresponds to a flat prior on γ*. Although the prior* $\pi(\theta) = \theta^{-1}(1-\theta)^{-1}\mathbb{I}_{(0,1)}(\theta)$ *may work well in the sense that apparently the associated posterior will not depend on hyperparameters, one should never forget that it implies that* θ *a priori is concentrated around* 0 *or* 1*. This is really a weird prior choice.*

The moral of this story is that a flat prior in one parameterization is not necessarily flat in another. So there is nothing intrinsically non-informative about a flat prior, at least according to what the negation of the English definition of the word *informative* would imply.

1.6.2 Jeffreys Priors

Jeffreys proposed a class of non-subjective priors for Bayesian problems that can often be termed reference priors. A feature of the Jeffreys prior is that it is often improper. However, it has been found to be useful in a number of instances, so it is worth to discuss it. The Jeffreys prior is defined as

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta)$ is the **Fisher (expected) information**:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(y|\theta)\right)^2\right] = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log f(y|\theta)\right]$$

where the expectation above is the conditional distribution of the data Y. As mentioned above, this quantity, as a function of θ, will often integrate to infinity, resulting in an improper prior specification. Jeffreys proposed such prior because it is invariant to monotone transformations, meaning that if we consider a reparameterization of the model, say $\gamma = g(\theta)$ where g is monotone over the domain of θ, then the Jeffreys prior for γ is precisely the same as the induced prior that we would obtain for γ using the transformation formula $\pi_\gamma(\gamma) = \pi_\theta\left(g^{-1}(\gamma)\right)\left|\frac{d}{d\theta} g^{-1}(\gamma)\right|$.

Example 1.4 (Ex. 6.4, Rosner et al.)**.** *Suppose we have independent normal observations with known mean* μ_0 *and unknown variance* $\frac{1}{\tau}$*,* $Y_1, \ldots, Y_n|\sigma^2 \overset{iid}{\sim} \mathcal{N}\left(\mu_0, \sigma^2\right)$*. The likelihood of* $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})$*, as a function of* τ*, is given by:*

$$f(\mathbf{y}|\tau) \propto \prod_{i=1}^{n} \tau^{\frac{1}{2}} \mathrm{e}^{-\frac{\tau}{2}(y_i-\mu_0)^2} = \tau^{\frac{n}{2}} \mathrm{e}^{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu_0)^2}$$

In this case, the log-likelihood is

$$\log f(y|\tau) \propto \frac{n}{2}\log\tau + \left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu_0)^2\right)$$

so that

$$\frac{d^2}{d\tau^2}\log f(y|\tau) = -\frac{n}{2}\tau^{-2}$$

Taking the expectation wrt the distribution of the data does not change the value of this constant, so that the Jeffreys prior is given by $\pi(\tau) \propto \sqrt{\frac{n}{2}\tau^{-2}} \propto \frac{1}{\tau}$ *if* $\tau > 0$.

1.6.3 Scientifically Informed Priors

It is important that we elicit scientifically relevant information for parameters that are scientifically relevant, as we mentioned above. For instance, if θ represents the probability of infection in a population of interest then its relevance is automatic and information should be gathered from external studies in order to set the prior $\pi(\theta)$.

Example 1.5. *Consider two independent binomial samples:* $Y_i|\theta_i \sim Bin(n_i, \theta_i)$, $\theta_i \in (0,1)$, $i = 1, 2$. *For instance,* θ_1 *might be the proportion of smokers that developed lung cancer during a fixed period of time and* θ_2 *might be the corresponding proportion of non-smokers. Both* θ_1 *and* θ_2 *are scientifically relevant and the interest is on comparing such proportions by looking at* $\theta_1 - \theta_2$ *or* $\frac{\theta_1}{\theta_2}$. *A natural choice for the priors would be to choose* θ_1 *and* θ_2 *independent and beta distributed so that the posteriors are still beta distributions. Therefore we could choose* $\theta_i \overset{ind}{\sim} beta(a_i, b_i), i = 1, 2$ *where the hyperparameters* (a_1, b_1) *and* (a_2, b_2) *are to be chosen according to available information on the proportions of lung cancer patients among smokers and non-smokers. We could also consider a different parametrization, typically for computational convenience. One such parametrization is the following:*

$$\theta_1 = \frac{e^{\beta_1}}{1+e^{\beta_1}}, \quad \theta_2 = \frac{e^{\beta_1+\beta_2}}{1+e^{\beta_1+\beta_2}}$$

so that

$$\beta_1 = \log\frac{\theta_1}{1-\theta_1}, \quad \beta_2 = \log\frac{\frac{\theta_2}{1-\theta_2}}{\frac{\theta_1}{1-\theta_1}} \text{ with } \beta_1, \beta_2 \in \mathbb{R}.$$

One possible choice for the prior of β_1, β_2, *simple from the computational viewpoint, would be to set* $\beta_i \overset{ind}{\sim} \mathcal{N}(\mu_{0i}, \sigma_{0i}^2)$, $i = 1, 2$. *However, choosing values for the hyperparameters that are "reasonable" with respect to the available information is not immediate in this case. Nonetheless, one should always check that for a given choice of such parameters the induced prior on the scientifically relevant parameters* θ_1 *and* θ_2 *is reasonable. For instance, see the plot of the marginal prior distributions induced on* θ_1 *and* θ_2 *when* $\mu_{01} = \mu_{01} = 0$ *and* $\sigma_{01}^2 = 10$, $\sigma_{01}^2 = 100$, *corresponding to vague marginal priors for* β_1, β_2.

The take-home message of this example is that prior information should be elicited for scientifically relevant parameters both directly and indirectly, no matter how little information may be available. In particular, if there are many parameters we need to be informative about the subset of parameters which are scientifically relevant and we also want to choose the priors for the remaining parameters so as to have a small impact on the analysis. Moreover, very often we will choose the form of the marginal priors in a convenient way from the algorithmic point of view.

1.6.4 Merging of the Priors

Two different experts with distinct subjective prior belief would choose two different prior distributions which will consequently lead to different posterior distributions. In this case one may ask how to

establish which prior corresponds to the best choice. While an answer for such question is hard to find, it is still possible to state that if the sample size n is large enough, then the posterior distributions will merge: the total variation distance between the two probability distributions converges to 0 as $n \to +\infty$. The reason behind such a phenomenon is because of the **consistency** of the posterior distribution: if θ_0 denotes the true value of θ then the posterior will concentrate more and more mass in a neighborhood of θ_0 as $n \to +\infty$.

1.7 Asymptotic Normality of the Posterior Distribution

Theorem 1.5. *Let $Y_1, \ldots, Y_n|\theta \overset{iid}{\sim} f(\cdot|\theta)$ with $\theta \sim \pi(\theta), \theta \in \Theta \subset \mathbb{R}^p$. Under suitable regularity conditions and for n large we have that the posterior distribution $\pi(\theta|\boldsymbol{y})$ can be approximated by any of the following distributions:*

1. $N_p\left(\widetilde{\theta}_n, V_n\right)$.
2. $N_p\left(\overline{\theta}_n, \left(\widetilde{I}_n\right)^{-1}\right)$.
3. $N_p\left(\widehat{\theta}_n, \left(\widehat{I}_n\right)^{-1}\right)$.
4. $N_p\left(\widehat{\theta}_n, \left(I_n\left(\widehat{\theta}_n\right)\right)^{-1}\right)$.

where:

- $\widetilde{\theta}_n$ *is the posterior mean.*
- $\overline{\theta}_n$ *is the posterior mode.*
- $\widehat{\theta}_n$ *is the maximum likelihood estimator of θ.*
- V_n *is the posterior covariance matrix.*
- $\widetilde{I}_n = \left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \pi(\theta|\boldsymbol{y})|_{\overline{\theta}_n}\right]_{ij}$ *is the generalized observed Fisher information matrix.*
- $\widehat{I}_n = \left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f(\boldsymbol{y}|\theta)|_{\widehat{\theta}_n}\right]_{ij}$ *is the observed Fisher information matrix.*
- $I_n(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f(\boldsymbol{y}|\theta)\right]_{ij}$ *is the (expected) Fisher information matrix.*

2 Simulation Methods for Bayesian Statistics

2.1 The Monte Carlo Method

Let us consider a random variable $\theta \in \Theta \subset \mathbb{R}^d$ and suppose that θ is distributed according to a target distribution π. In our applications, π is always the posterior of the Bayesian framework. The Monte Carlo method is used to learn relevant features of the target distribution π through simulation. In particular, the method assumes that it is possible to draw independent and identically distributed samples from π: $\theta^{(1)}, \theta^{(2)}, \ldots \overset{\text{iid}}{\sim} \pi$, we need to simulate a large (infinite) number of draws from the target distribution π.

Theorem 2.1 (Strong Law of Large Numbers)**.** *Let π be a distribution over Θ and $h : \Theta \to \mathbb{R}$ be a function such that $\int_\Theta |h(\theta)| \pi(d\theta) < +\infty$. Moreover, let $\theta^{(1)}, \theta^{(2)}, \ldots$ be a sequence of random variables independent and identically distributed from π. Then*

1. *The sequence of random variables $\overline{h}^{(T)} := \frac{1}{T} \sum_{i=1}^{T} h\left(\theta^{(i)}\right) \xrightarrow[T \to +\infty]{a.s.} \int_\Theta h(\theta) \pi(d\theta) = \overline{h}$.*
2. *For any $p \in (0,1)$, if q_p is the p-quantile of $h(\theta)$ and $q_p^{(T)}$ is the empirical p-quantile of $\{h\left(\theta^{(1)}\right), \ldots, h\left(\theta^{(T)}\right)\}$, then $q_p^{(T)} \xrightarrow[T \to +\infty]{a.s.} q_p$.*

The SLLN is used to implement the Monte Carlo method: indeed, the features of the distribution π can be approximated by computing $\overline{h}^{(T)}$ for a smart choice of the function h. For instance, if we are interested in approximating the quantity $\pi(A)$, this quantity is expressed as $\bar{h}$, with $h(\theta) = \mathbb{I}_A(\theta)$. If we are interested in approximating $\bar{h} = E_\pi[\theta_i]$, where θ_i is the i-th component of θ, in this case the function h is the projection $h(\theta) = \theta_i$. Observe, however, that while the SLLN guarantees that $\text{err}^{(T)} = \overline{h}^{(T)} - \overline{h} \xrightarrow[T \to +\infty]{a.s.} 0$, such theorem does not say anything regarding the speed of the convergence. For such a job we need to resort to the following theorem.

Theorem 2.2 (Central Limit Theorem)**.** *Let π be a distribution over Θ and $h : \Theta \to \mathbb{R}$ be a function such that $\int_\Theta |h(\theta)| \pi(d\theta) < +\infty$ and $0 < \text{Var}(h(\theta)) =: \sigma^2 < +\infty$. Moreover, let $\theta^{(1)}, \theta^{(2)}, \ldots$ be a sequence of random variables independent and identically distributed from π We have that:*

1. $\sqrt{T}\left(\overline{h}^{(T)} - \overline{h}\right) \xrightarrow[T \to +\infty]{d} \mathcal{N}\left(0, \sigma^2\right)$.
2. $\sigma^{2(T)} = \frac{1}{T} \sum_{i=1}^{T} \left(h\left(\theta^{(i)}\right) - \overline{h}^{(T)}\right)^2 \xrightarrow[T \to +\infty]{a.s.} \sigma^2$.

In particular, the CLT tells us that if T is large then $\text{err}^{(T)} = \overline{h}^{(T)} - \overline{h} \approx \mathcal{N}\left(0, \frac{\sigma^{2(T)}}{T}\right)$ so that, for any $c > 0$:

$$\mathbb{P}\left(\left|\text{err}^{(T)}\right| > c\right) = \mathbb{P}\left(\frac{\left|\text{err}^{(T)}\right|}{\sqrt{\frac{\sigma^{2(T)}}{T}}} > \frac{c}{\sqrt{\frac{\sigma^{2(T)}}{T}}}\right) \simeq 2\left(1 - \Phi\left(\frac{c}{\sqrt{\frac{\sigma^{2(T)}}{T}}}\right)\right)$$

If T is large, then $\sqrt{\frac{\sigma^{2(T)}}{T}}$ is small, and $\frac{c}{\sqrt{\frac{\sigma^{2(T)}}{T}}}$ is large, so that the probability that the absolute value of $err^{(T)}$ exceeds c is close to 0. This imply that $\bar{h}^{(T)}$ is a good estimator of $\bar{h}$.

Example 2.1 (see Hoff (2009), Chapter 4)**.** *Over the course of the 1990s, the General Social Survey gathered data on the educational attainment and number of children of 155 women. In this example we will compare the women with college degrees to those without in terms of their numbers of children. Let $Y_{11}, \ldots, Y_{n_1 1}$ denote the numbers of children for the $n_1 = 111$ women without college degrees and $Y_{12}, \ldots, Y_{n_2 2}$ be the data for the $n_2 = 44$ women with degrees. We will assume the following model:*

$$Y_{11}, \ldots, Y_{n_1 1} | \theta_1 \overset{iid}{\sim} Poisson(\theta_1)$$

and

$$Y_{12}, \ldots, Y_{n_2 2} \overset{iid}{\sim} Poisson\left(\theta_2\right)$$

with $\mathbf{Y_1} = (\mathbf{Y_{11}}, \ldots, \mathbf{Y_{n_1 1}})$ *and* $\mathbf{Y_2} = (\mathbf{Y_{12}}, \ldots, \mathbf{Y_{n_2 2}})$ *independent, conditionally to* (θ_1, θ_2)*. We assign the following prior:*

$$\theta_j \overset{iid}{\sim} gamma\left(a, b\right), j = 1, 2$$

with $a = 2, b = 1$*. It is straightforward to see that a posteriori* θ_1 *and* θ_2 *are still independent, and that, because the gamma is conjugate to the Poisson likelihood,*

$$(\theta_1, \theta_2)|\mathbf{y_1}, \mathbf{y_2} \sim gamma(\alpha + \sum_{\mathbf{i}} \mathbf{y_{i1}}, \beta + \mathbf{n_1}) \times gamma(\alpha + \sum_{\mathbf{i}} \mathbf{y_{i2}}, \beta + \mathbf{n_2}) = \mathbf{gamma(219, 112) \times gamma(68, 45)}$$

We are interested in computing $\mathbb{P}\left(\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2\right)$*.*
The true value of $\mathbb{P}\left(\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2\right)$ *(computed via numerical methods for integration) is 0.97. We approximate it via the Monte Carlo method, where* $h\left(\theta_1, \theta_2\right) = \mathbb{I}_{(\theta_1 > \theta_2)}$*. In particular, for* $T = 1000$ *we get* $\overline{h}^{(T)} = \frac{1}{T} \sum_{i=1}^{T} \mathbb{I}_{(\theta_1 > \theta_2)} \left(\theta_1^{(i)}, \theta_2^{(i)}\right) \approx 0.97$ *so that the estimate is accurate.*

2.1.1 Monte Carlo Method for the Posterior Predictive Distribution: The Augmentation Trick

We have seen that in the Bayesian framework, an object of interest is also the posterior predictive distribution (density in this case): $m_{Y_{n+1}|Y_1,\ldots,Y_n}\left(y|y_1, \ldots, y_n\right)$ which, in case of conditionally iid data, can be written as

$$m_{Y_{n+1}|Y_1,\ldots,Y_n}\left(y|y_1, \ldots, y_n\right) = \int_{\Theta} f\left(y|\theta\right) \pi\left(\mathrm{d}\theta|y_1, \ldots, y_n\right)$$

If we only need to evaluate this posterior predictive density via the Monte Carlo method, then $m_{Y_{n+1}|Y_1,\ldots,Y_n}\left(y|y_1, \ldots, y_n\right) \approx \frac{1}{T} \sum_{i=1}^{T} f\left(y\middle|\theta^{(i)}\right)$ for all y, where $f\left(\cdot|\theta\right)$ is the conditional density of the Y_i's and $\theta^{(1)}, \ldots, \theta^{(T)} \overset{\text{iid}}{\sim} \pi\left(\theta|y_1, \ldots, y_n\right)$. However, if we want to have an iid sample from the posterior predictive distribution, we need to apply the so-called **augmentation trick**.

In general, let X_1 be a random variable (element) such that we are not able to simulate directly from its distribution $\mathcal{L}\left(X_1\right)$. However, there exists X_2, another random variable (element), not necessarily of the same dimension as X_1, such that we can iid simulate from $\mathcal{L}\left(X_1|X_2\right)$ and $\mathcal{L}\left(X_2\right)$. In other words, we are able to simulate iid draws from the joint distribution $\mathcal{L}\left(X_1, X_2\right) = \mathcal{L}\left(X_1|X_2\right)\mathcal{L}\left(X_2\right)$. A sequence of independent and identically distributed draws of the law $\mathcal{L}\left(X_1\right)$ can be obtained through the following algorithm:

- For $t = 1, \ldots, T$ do the following:
  - Draw a sample $x_2^{(t)} \sim \mathcal{L}\left(X_2\right)$.
  - Draw a sample $x_1^{(t)} \sim \mathcal{L}\left(X_1\middle|X_2 = x_2^{(t)}\right)$ and append it to a list.

Indeed, since $\left(x_1^{(1)}, x_2^{(1)}\right), \ldots, \left(x_1^{(T)}, x_2^{(T)}\right) \overset{\text{iid}}{\sim} \mathcal{L}\left(X_1, X_2\right)$, then we have that $x_1^{(1)}, \ldots, x_1^{(M)} \overset{\text{iid}}{\sim} \mathcal{L}\left(X_1\right)$. In particular, since we have $m_{Y_{n+1}|Y_1,\ldots,Y_n}\left(y|y_1, \ldots, y_n\right) = \int_{\Theta} f\left(y|\theta\right) \pi\left(\mathrm{d}\theta|y_1, \ldots, y_n\right)$, we can simulate from the posterior predictive distribution by applying the augmentation trick where we set:

- $\mathcal{L}\left(X_1\right) = \mathcal{L}\left(Y_{n+1}|Y_1, \ldots, Y_n\right)$.
- $\mathcal{L}\left(X_1|X_2\right) = \mathcal{L}\left(Y_{n+1}|\theta\right)$ (which law is represented by the conditional density $f\left(\cdot|\theta\right)$).
- $\mathcal{L}\left(X_2\right) = \mathcal{L}\left(\theta|Y_1, \ldots, Y_n\right)$ (which law is represented by the density $\pi\left(\theta|Y_1, \ldots, Y_n\right)$).

2.2 Rejection Sampling

Rejection Sampling is a method to get a draw from a density we cannot directly sample from, by using another density that is easily sampled from. Let $\pi(\theta)$ be the target density and assume that we can evaluate $\pi(\theta)$ for any $\theta \in \Theta$. Moreover, suppose that there exist $c > 0$ and a density $m(\theta)$ such that we can sample from $m(\theta)$ and the inequality $\pi(\theta) < c \cdot m(\theta)$ holds. Note that if this equality hold, c must be larger than 1. Then the following algorithm generates the desired samples from the distribution with density $\pi(\theta)$:

- Draw a sample $z \sim m(\cdot)$ and compute $r(z) = \frac{\pi(z)}{c \cdot m(z)} \in (0, 1)$.
- Generate $u \sim \mathcal{U}(0, 1)$.
- If $u \leq r(z)$ accept $\widetilde{\theta} = z$, otherwise reject it and go back to the first step.

Let us prove that this procedure guarantees a sample from π (in the unidimensional case, i.e., θ is unidimensional.

Proof. We have that

$$\mathbb{P}\left(\widetilde{\theta} \leq t\right) = \mathbb{P}\left(Z \leq t \middle| U \leq \frac{\pi(Z)}{c \cdot m(Z)}\right) = \frac{\mathbb{P}\left(Z \leq t, U \leq \frac{\pi(Z)}{c \cdot m(Z)}\right)}{\mathbb{P}\left(U \leq \frac{\pi(Z)}{c \cdot m(Z)}\right)}$$

$$= \frac{\int_{\mathbb{R}} \mathbb{P}\left(Z \leq t, U \leq \frac{\pi(Z)}{c \cdot m(Z)} \middle| Z = z\right) m(z)\, \mathrm{dz}}{\int_{\mathbb{R}} \mathbb{P}\left(U \leq \frac{\pi(Z)}{c \cdot m(Z)} \middle| Z = z\right) m(z)\, \mathrm{dz}} = \frac{\int_{\mathbb{R}} \frac{\pi(z)}{c \cdot m(z)} \mathbb{I}_{(z \leq t)} m(z)\, \mathrm{dz}}{\int_{\mathbb{R}} \frac{\pi(z)}{c \cdot m(z)} m(z)\, \mathrm{dz}} = \int_{-\infty}^{t} \pi(z)\, \mathrm{dz}.$$

□

2.3 Markov Chain Monte Carlo Methods

The goal is to compute draws that are approximately from $\pi(\theta|\mathbf{y})$, the posterior distribution (density) of θ in our Bayesian model. In fact, only for very simple models we will be able to sample iid from the posterior. We will be able to sample draws of θ that are **approximately** from the posterior distribution.

2.3.1 General State Space Markov Chains

We now revive some theory for Markov chains with general state space (not necessarily denumerable). To exemplify, suppose the state space E to be a (measurable) subset of $\mathbb{R}$ of $\mathbb{R}^k$ for some integer k.

Definition 2.1 (Time-homogeneous Markov chain)**.** *Let $\{X_n, n \geq 0\}$ be a sequence of random elements with values in a set $E \subset \mathbb{R}^k$. We say that $\{X_n, n \geq 0\}$ is a* ***time-homogeneous Markov chain with state space*** **E** *if*

$$\mathbb{P}(X_{n+1} \in A | X_0 = x_0, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} \in A | X_n = x_n) =: P(x_n, A),$$

for all $A \in \mathcal{E}$, all $x_0, x_1, \ldots, x_n \in E$ and all n.

Definition 2.2 (Transition probability kernel)**.** *The function $P(x_n, A) := \mathbb{P}(X_{n+1} \in A | X_n = x_n)$ is called* ***transition probability kernel*** *and is such that:*

- *$x \mapsto P(x, A)$ is measurable for all $A \in \mathcal{E}$.*
- *$A \mapsto P(x, A)$ is a probability on $(E, \mathcal{E})$ for all $x \in E$.*

*Similarly, we define the n-**step transition probability kernel** the object $P^n(x, A) := \mathbb{P}(X_n \in A | X_0 = x)$.*

Remark. *It is possible to compute the n-step transition probability kernel $P^n(x,\cdot)$ iteratively through the Chapman-Kolmogorov equation:*

$$n = 1 : P^1(x, A) = P(x, A)$$

$$n = 2 : P^2(x, A) = P(X_2 \in A | X_0 = x) = \int_E P(x, dy) P(y, A)$$

$$\vdots$$

$$\text{for any } n \geq 2 : P^n(x, A) = \int_E P(x, dy) P^{n-1}(y, A)$$

Definition 2.3 (Invariant probability measure)**.** *A probability measure π on $(E, \mathcal{E})$ is called **invariant** (or **stationary**) for the Markov chain $\{X_n, n \geq 0\}$ if we have that*

$$\int_E P(x, A) \pi(dx) = \pi(A) \quad \text{for all } A \in \mathcal{E}.$$

In this case we write $\pi P = \pi$.

Definition 2.4 (Irreducibility)**.** *Let ϕ be a probability measure on $(E, \mathcal{E})$. We say that $\{X_n, n \geq 0\}$ is **ϕ-irreducible** if for any $A \in \mathcal{E}$ such that $\phi(A) > 0$ there exists $n = n(x, A) \geq 1$ with $P^n(x, A) > 0$ for all $x \in E$. Moreover, we say that $\{X_n, n \geq 0\}$ is **irreducible** if there exists a probability ϕ (which is called **irreducibility distribution**) such that the chain is ϕ-irreducible.*

Remark. *If the chain is irreducible then it will eventually visit all the remarkable states. This will imply that the MCMC visits the whole support of the posterior distribution.*

The following theorem will apply for Metropolis-Hastings Markov chains and for the MC defined by the Gibbs sampling algorithm.

Theorem 2.3. *We have that:*

- *A Markov chain with transition probability kernel $P(x, A)$ is ϕ-irreducible if there exists $n \geq 1$ such that P^n has a (strictly) positive density f (with respect to ϕ).*
- *If the transition probability kernel P has discrete and absolutely continuous components and there exists $n \geq 1$ such that the continuous component of P^n has a (strictly) positive density with respect to ϕ then P is ϕ-irreducible.*

Remark. *If $\{X_n, n \geq 0\}$ is irreducible then there exist different irreducibility distributions. However, they are all absolutely continuous with respect to one of them, which is called **maximal irreducibility distribution** and is denoted by Ψ^*.*

If irreducibility means that all *remarkable* sets may be reached (sooner or later), the *recurrence* means that these sets may be reached *i.o.* (infinitely often), at least from almost every initial point. Note that X_n visits the set A *infinitely often* if

$$X_n \in A \text{ i.o. } \Leftrightarrow \forall n \; \exists k \geq n : X_k \in A \Leftrightarrow \cap_{n=1}^{+\infty} \cup_{k \geq n} \{X_k \in A\}.$$

Definition 2.5 (Recurrency)**.** *Let $\{X_n, n \geq 0\}$ be an irreducible Markov chain with maximal irreducibility distribution Ψ^*. Then $\{X_n, n \geq 0\}$ is called **recurrent** if for all $A \in \mathcal{E}$ such that $\Psi^*(A) > 0$ we have that:*

- $\mathbb{P}_x(X_n \in A \text{ i.o. }) > 0$ *for all x.*

- $\mathbb{P}_x(X_n \in A \text{ i.o. }) = 1$ *almost everywhere with respect to* x*, i.e. the measure* Ψ^* *of all* x *for which the equality does not hold is equal to 0.*

We have introduced this new notation: $\mathbb{P}(\cdot|X_0 = x)$ with $\mathbb{P}_x(\cdot)$.

Remark. *In a nutshell, we have that, if the chain is recurrent, then every remarkable set* A *is visited infinitely often by the chain with probability equal to* 1 *starting from almost any point* $x \in E$*. This will guarantee that the posterior probability (if positive) assigned to any subset* A *can be approximated by the relative frequency of visits to* A *of the MCMC.*

Definition 2.6 (Positive recurrency)**.** *An irreducible and recurrent Markov chain is called* ***positive recurrent*** *if it admits an invariant probability, otherwise the chain is called* ***null****.*

Theorem 2.4. *Let* $\{X_n, n \geq 0\}$ *be irreducible and let* π *be a stationary distribution. Then we have that:*

- $\{X_n, n \geq 0\}$ *is* π*-irreducible and* π *is the maximal irreducibility distribution.*
- π *is the unique stationary distribution.*
- $\{X_n, n \geq 0\}$ *is positive recurrent.*

Theorem 2.5 (Ergodic Theorem)**.** *Let* $\{X_n, n \geq 0\}$ *be an irreducible Markov chain and let* π *be its (unique) invariant distribution. Moreover, consider a function* $h : E \to \mathbb{R}$ *such that* $\mathbb{E}_\pi[|h|] := \int |h(x)| \pi(dx) < +\infty$*. Then we have that:*

$$\mathbb{P}_x\left(\frac{1}{n+1}\sum_{i=0}^{n} h(X_i) \to \int_E h d\pi\right) = 1 \quad \pi - a.e. \text{ wrt } x.$$

Remark. *The limit may not hold for initial points* x *belonging to a set* C *such that* $\pi(C) = 0$*.*

How will we typically use this theorem? Suppose now that we are interested in computing $\pi(A)$ for some target distribution π. Since $\pi(A) = \int h d\pi$ with $h(x) = \mathbb{I}_A(x)$, we can use the Ergodic Theorem to approximate $\pi(A)$ by constructing an irreducible Markov chain with invariant distribution equal to π and approximating $\pi(A)$ with

$$\widehat{\pi}(A) := \frac{|\{i : x_i \in A\}|}{m+1}$$

for a given realization of the chain $x_0, \ldots, x_m$ with m large. However, this is true only if the choice for the initial condition x_0 is "lucky", meaning that it does not belong to the zero-measure set C of initial conditions for which the Ergodic Theorem does not hold. Therefore, it is preferable to achieve a stronger result (i.e. the thesis holds for all $x \in E$), though assuming a stronger condition than recurrence.

Definition 2.7 (Harris-recurrent Markov chain)**.** *Let* $\{X_n, n \geq 0\}$ *be an irreducible Markov chain with maximal irreducibility distribution* Ψ^**. We say that* $\{X_n, n \geq 0\}$ *is* ***Harris-recurrent*** *if for all* A *such that* $\Psi^*(A) > 0$ *we have that* $\mathbb{P}_x(X_n \in A \text{ i.o. }) = 1$ *for all* $x \in E$*.*

Theorem 2.6 (Ergodic Theorem for Harris-recurrent Markov Chains)**.** *Let $\{X_n, n \geq 0\}$ be an irreducible and Harris-recurrent Markov chain with invariant distribution π and let $h : E \to \mathbb{R}$ be such that $\mathbb{E}_\pi [|h|] < +\infty$. Then we have that:*

$$\mathbb{P}_x \left(\frac{1}{n+1} \sum_{i=0}^{n} h(X_i) \to \int_E h d\pi \right) = 1 \text{ for all initial point } x \in E.$$

Note also that, if the chain is aperiodic (see definition below), then we have a stronger convergence result:

Theorem 2.7. *Let $\{X_n, n \geq 0\}$ be an aperiodic, irreducible and Harris-recurrent Markov chain with invariant distribution π. Then*

$$\|P^n(x, \cdot) - \pi(\cdot)\| \overset{n \to +\infty}{\longrightarrow} 0 \text{ for all } x \in E.$$

In practice, the theorem above means that, if the Markov chain is also Harris-recurrent, then $\mathcal{L}(X_n | X_0 = x) \approx \pi$ for n large enough and any initial point x.

Definition 2.8. *Let $\{X_n, n \geq 0\}$ be irreducible; a m-cycle is a family of disjoint sets $\{E_0, E_1, \ldots, E_{m-1}\}$ such that*

$$P(x, E_j) = 1 \quad j = i + 1 \mod m \quad \text{for all } x \in E_i.$$

The period of the chain d is the largest integer m for which an m-cycle exists. When $d = 1$, the chain is called aperiodic.

Note that MCs that are aperiodic, irreducible and Harris-recurrent are called *Harris-ergodic*.

We can now finally describe how Markov Chain Monte Carlo methods work. As in the case of the Monte Carlo method, the goal is to compute $\mathbb{E}_\pi [h(\theta)]$ where π is some target distribution (i.e. the posterior distribution in the Bayesian framework) and $h : \Theta \to \mathbb{R}$ is such that the integral exists and is finite. Using the previous results it is easy to realize that a suitable procedure to do that is the following:

- Build a Markov chain $\{\theta_n, n \geq 0\}$ with state space Θ which is irreducible and Harris-recurrent and such that its invariant distribution is given by π.
- Select an initial condition θ_0 by choosing a value that is reasonable according to the scientific interpretation and knowledge about the parameter θ.
- Simulate the Markov chain. In particular, for a sufficiently large number BI (which is called **burn-in**) we have that $\theta_t \sim \pi$ at least approximately, if $t > BI$, thanks to the Ergodic theorem.
- Approximate the quantity $\mathbb{E}_\pi [h(\theta)]$ with the quantity $\frac{1}{T - BI} \sum_{i=BI+1}^{T} h(\theta_i)$ for T large.

Remark. *The Ergodic Theorem 2.6 allows us to approximate $\mathbb{E}_\pi [h(\theta)]$ with the whole ergodic mean $\overline{h}_T = \frac{1}{T+1} \sum_{i=0}^{T} h(\theta_i)$. However, it is convenient to consider only the iterates that come after the burn-in threshold for efficiency reasons.*

We now need to discuss how we can check the assumptions of the Ergodic Theorem 2.6 in order to apply MCMC methods. For what regards irreducibility, we have already stated two criteria which will apply to the two core examples of MCMC algorithms (Gibbs sampler and Metropolis-Hastings algorithm). Moreover, it is possible to prove that such criteria also guarantee Harris-recurrence for the Markov chain. Hence we need only to build the chain such that π, the target distribution, is the invariant distribution of the chain.
Let us introduce the notion of reversibility to deal with the choice of the invariant distribution for the chain.

Definition 2.9 (Reversibility). *Let $\{X_n, n \geq 0\}$ be a Markov chain with transition probability kernel P. We say that the chain is **reversible** with respect to a probability distribution π on E if we have that $\pi(dx) P(x, dy) = \pi(dy) P(y, dx) \forall x, y \in E$.*

Remark. *If π and P have densities then reversibility is equivalent to*

$$\pi(x) p(x, y) = \pi(y) p(y, x) \text{ for all } x, y \in E.$$

Theorem 2.8. *If $\{X_n, n \geq 0\}$ is reversible with respect to the probability π then π is invariant for the chain.*

Proof. We have that:

$$\int_{x \in E} \pi(dx) P(x, dy) = \int_{x \in E} \pi(\mathrm{dy}) P(y, dx) = \pi(dy)$$

□

Does the Central Limit Theorem for Markov chains hold? We have seen that the law of large numbers does hold. Let us introduce two more definitions.

Definition 2.10 (Geometrically and uniformly ergodic chains). *Let $\{X_n, n \geq 0\}$ be Harris-ergodic and let π be its invariant distribution. The chain is called **geometrically ergodic** if there exist $M : E \to \mathbb{R}^+$ with $\mathbb{E}_\pi[M] < +\infty$ and $r \in (0, 1)$ such that $\|P^n(x, \cdot) - \pi(\cdot)\| \leq M(x) r^n$ for all $x \in E$ and all n. Moreover, if $M(x) = const$, the chain is called **uniformly ergodic**.*

Note that necessary condition to assume geometric or uniform ergodicity is that the chain is Harris ergodic, i.e. Theorem 2.6 holds.

Theorem 2.9 (Central Limit Theorem). *Let $\{X_n, n \geq 0\}$ be an Harris-ergodic Markov chain and let π be its invariant distribution. Moreover, let $h : E \to \mathbb{R}$ such that one of these conditions holds:*

a) $\{X_n, n \geq 0\}$ is geometrically ergodic and $\mathbb{E}_\pi\left[|h|^{2+\varepsilon}\right] < +\infty$ for some $\varepsilon > 0$.

b) $\{X_n, n \geq 0\}$ is uniformly ergodic and $\mathbb{E}_\pi\left[|h|^2\right] < +\infty$.

Then we have that:

$$\sqrt{n}\left(\overline{h}_n - \int h d\pi\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_h^2\right) \text{ as } n \to +\infty$$

where $\bar{h}_n := \frac{1}{n+1} \sum_{i=0}^{n} h(X_i)$ and

$$\sigma_h^2 = Var_\pi(h(X_0)) + 2 \sum_{k=1}^{+\infty} Cov_\pi(h(X_0), h(X_k)) \text{ with } X_k \sim \pi \quad k = 0, 1, 2, \ldots. \tag{6}$$

Remark. *Observe that the variance of the estimator $\bar{h}_n$ of $\int h d\pi$ is given by $\frac{\sigma_h^2}{n}$ and it is larger (in general) than the variance in the case of the Central Limit Theorem for iid variables, because in this case here the iterates are not independent. Due to this fact, we have that the number of samples after the burn-in threshold ($T - BI$) needs to be larger than the number of samples we would use for the Monte Carlo method.*

For more details on the theory on general state space Markov chains, see Tierney (1994) or Jackman (2009) (Ch. 4).

2.4 The Metropolis-Hastings Algorithm

Let us assume that the target distribution π has a density with respect to some measure μ (e.g., the Lebesgue measure) and let us call $\pi(x)$ such density. Moreover, set $E^+ := \{x \in E : \pi(x) > 0\}$ and consider a transition probability kernel $Q(x, dy) = q(x, y)\,\mu(\mathrm{dy})$ such that $Q(x, E^+) = 1$ for all $x \notin E^+$ (i.e. if we are not in E^+ at a certain time we will reach E^+ at the next iteration). The pseudo-code of the Metropolis-Hastings algorithm is the following:

1. At iteration n, let the chain be in state $X_n = x$; then generate a candidate y from the distribution $Q(x, \cdot)$.
2. Set $X_{n+1} = y$ with acceptance probability $\alpha(x, y) := \begin{cases} \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) & \text{if } \pi(x)\,q(x, y) > 0 \\ 1 & \text{if } \pi(x)\,q(x, y) = 0 \end{cases}$, otherwise set $X_{n+1} = x$.
3. Advance the iteration index to $n + 1$ and go to 1.

Remark. *Observe that in order to build a Metropolis-Hastings chain we need to:*

- *be able to sample from the density $q(x, \cdot)$ for every $x \in E$.*
- *be able to evaluate the two ratios $\frac{\pi(y)}{\pi(x)}$ and $\frac{q(y,x)}{q(x,y)}$.*

If we define

$$p(x, y) := \begin{cases} q(x, y)\,\alpha(x, y) & x \neq y \\ 0 & x = y \end{cases},$$

then the transition probability kernel of the Markov chain resulting from this procedure is the following:

$$P(x, dy) = p(x, y)\,\mu(\mathrm{dy}) + r(x)\,\delta_x(\mathrm{dy})$$

that is equivalent to

$$P(x, A) = \int_A p(x, y)\,\mu(\mathrm{dy}) + r(x)\,I_A(x). \tag{7}$$

The expression $r(x) = \mathbb{P}(X_{n+1} = x | X_n = x)$ can be recovered from (7) from $A = E$, that is

$$1 = P(x, E) = \int_E p(x, y)\,\mu(\mathrm{dy}) + r(x)\,I_E(x),$$

so that $r(x) = 1 - \int_E q(x, y)\,\alpha(x, y)\,\mu(\mathrm{dy})$.
Let us now check that π is a stationary distribution for the chain. Since stationarity is equivalent to reversibility, it suffices to show that:

$$\pi(x)\,p(x, y) = \pi(y)\,p(y, x) \text{ for all } x, y, \text{ i.e. } q(x, y)\,\alpha(x, y)\,\pi(x) = \pi(y)\,q(y, x)\,\alpha(y, x)$$

Assume, for instance, that $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \leq 1$ so that $\alpha(x, y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$. In this case we have that $LHS = q(x, y)\,\alpha(x, y)\,\pi(x) = q(x, y)\,\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\pi(x) = \pi(y)\,q(y, x) = RHS \iff \alpha(y, x) = 1$. Moreover, since $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \leq 1$ we have that $\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)} > 1 \Longrightarrow \alpha(y, x) = 1$. The same can be shown for the case $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} > 1$ through a similar procedure.
The second condition to check to have an ergodic MH chain is the irreducibility of the chain; in this case, we need stronger assumptions on the proposal density $q(x, y)$. In particular, the following options are the most popular choices:

- **Random walk Metropolis-Hastings chain:** Let $q(x, y) = f(y - x)$ where f is a density, which is equivalent to setting $Y = x + Z$ with $Z \sim f$ (typically $Z \sim \mathcal{N}$ or $Z \sim t$).
  In this case we have that if $f(x) > 0$ for all $x \in E$, then the chain $(X_n)_n$ is (π-)irreducible, recurrent and aperiodic.
  If the condition above is not valid but we have that E^+ is an open connected set and that $f(x) > 0$ for all $x \in \mathcal{U}(\mathbf{0})$ (where $\mathcal{U}(\mathbf{0})$ is a neighborhood of the origin of E), then the conclusion still holds.

- **Independence Metropolis-Hastings chain:** Let $q(x,y) = f(y)$ where f is a density that does not depend on x, which is equivalent to setting $Y \sim f$.
  In this case we have that if $f(x) > 0$ almost everywhere on E^+ with respect to μ, then the chain $(X_n)_n$ is (π-)irreducible, recurrent and aperiodic.

In general we have that for a Metropolis-Hastings chain the followings hold:

- If $r(x) > 0$ almost everywhere with respect to μ then the chain is aperiodic.
- If $q(x,y) > 0$ for all $x, y \in E^+$ then the chain is irreducible.
- If the chain is irreducible then it is Harris-recurrent.

All in all, in order to build a MH chain,

- we are required to sample from the density $q(x,\cdot)$
- we must be able to evaluate two ratios, $\dfrac{\pi(y)}{\pi(x)}$ and $\dfrac{q(y,x)}{q(x,y)}$

Remark. *For the MH algorithm, a candidate point y is accepted if $\dfrac{\pi(y)}{q(x,y)} > \dfrac{\pi(x)}{q(y,x)}$; in case of a RWMH, the candidate point y is accepted if $\pi(y) > \pi(x)$.*

Let us spend a few words on the acceptance rate for the candidates of the RWMH chain. In practice, it is recommended that such rate should be between 0.2 and 0.5. Indeed, a rate that is too low would result in a chain which is often "stuck" on the same value, which would require a large number of simulations in order to sweep the space E appropriately. On the other hand, a rate that is too high is associated with a high correlation of the iterates, resulting in the need for a large number of simulations as stated in the Central Limit Theorem. Such rate is typically controlled through a tuning parameter in the density $q(x,y)$.

For further reference, see Rosner et al. (2021), Section 4.4.2 and Jackman (2009).

Example 2.2 (see Albert (2009), Section 6.7)**.** *We have collected $n = 211$ measures of the heights (in inches) of male students in a campus, though data are grouped: We assume that the underlying*

Height	< 66	$[66,68)$	$[68,70)$	$[70,72)$	$[72,74)$	≥ 74
Frequency n_i	14	30	49	70	33	15

observations can be modeled as $Y_1, \dots, Y_n | \mu, \sigma \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. However, since data are grouped, the joint (conditional) distributions is $L(\mu, \sigma; \boldsymbol{y}) = \prod_{i=1}^{6} p_i^{n_i}$ where $p_i = \mathbb{P}(Y_i \in C_i)$. For instance $p_1 = P(Y_1 < 66) = \Phi(\dfrac{66-\mu}{\sigma})$.
The author assumes the following prior: $\pi(\mu, \sigma) \propto \frac{1}{\sigma}\mathbb{I}_{(0,+\infty)}(\sigma)$ (improper prior). Please, the use of improper priors is discouraged, and we only adopt here in this example. We now transform the standard deviation by $\lambda = \log\sigma$ so that the prior is such that $\pi(\mu, \lambda) \propto const$ and the posterior is $\pi(\mu, \lambda | \boldsymbol{y}) \propto L(\mu, \lambda, \boldsymbol{y})$.
Let us construct a random walk Metropolis-Hastings chain to simulate from the posterior.... Se the R notebook in WeBeep.

2.5 Gibbs Sampler

The Metropolis-Hastings algorithm is not very efficient when θ is multidimensional as it requires to sample from a multidimensional Gaussian (or Student's t) distribution. Moreover, the use of a joint proposal density is not appropriate if the components of the posterior are scaled quite differently or if there is multimodality or a pronounced skew along one component. It is therefore natural to look

for an alternative algorithm with a divide-and-conquer nature: such algorithm is given by the Gibbs sampler, which requires the sampling from univariate distributions.

The parameter vector is bidimensional

Let $\theta = (X, Y)$. Denote by $\pi(x, y)$ the target distribution and assume that we are able to sample from the full conditional distributions $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. The pseudo-code for the algorithm is the following:

1. At iteration n, let the chain be in state $(X_n = x_n, Y_n = y_n)$, then update the first component by simulating
$$X_{n+1} \sim f_{X|Y}(\cdot|y_n)$$
2. Update the second component by simulating
$$Y_{n+1} \sim f_{Y|X}(\cdot|x_{n+1})$$
3. Advance the iteration index to $n + 1$ and go to 1.

The output of such algorithm is a bivariate Markov chain $(X_n, Y_n)_{n\geq 0}$. We have:

Theorem 2.10. *π is the invariant distribution of $(X_n, Y_n)_{n\geq 0}$.*

Theorem 2.11. *If $\pi_X(x) > 0, \pi_Y(y) > 0$ imply $\pi(x, y) > 0$ (**positivity condition**), then $(X_n, Y_n)_{n\geq 0}$ is (π-)irreducible.*

Theorem 2.12. *If the transition probability kernel $P((x, y), A \times B) = \int_{A\times B} f_{X|Y}(x_1|y) f_{Y|X}(y_1|x_1) dx_1 dy_1$ is absolutely continuous with respect to the invariant distribution π then $(X_n, Y_n)_{n\geq 0}$ is Harris-recurrent.*

In a nutshell, we have that the Gibbs sampler works when:

- $\text{supp}(\pi_X) \times supp(\pi_Y) = supp(\pi)$, which is equivalent to the positivity condition.
- $f_{X|Y}(\cdot|y) > 0$ and $f_{Y|X}(\cdot|x) > 0$ on the support sets of the marginals π_X and π_y.
- The marginal π_X and π_Y exist (i.e. π is not an improper distribution).

The parameter vector is p-dimensional, $p > 2$

The pseudo-code for the algorithm is the following:

- At iteration n, let the chain be in state $\left(X_1^{(n)} = x_1^{(n)}, \ldots, X_p^{(n)} = x_p^{(n)}\right)$, then update the first component by simulating
$X_1^{(n+1)} \sim f_{X_1|X_2,\ldots,X_p}\left(\cdot\middle|x_2^{(n)}, \ldots, x_p^{(n)}\right)$.
- Update the second component by simulating
$X_2^{(n+1)} \sim f_{X_2|X_1,X_3,\ldots,X_p}\left(\cdot\middle|x_1^{(n+1)}, x_3^{(n)}, \ldots, x_p^{(n)}\right)$.
- $\vdots$
- Update the p-th component by simulating
$X_p^{(n+1)} \sim f_{X_p|X_1,\ldots,X_{p-1}}\left(\cdot\middle|x_1^{(n+1)}, \ldots, x_{p-1}^{(n+1)}\right)$.
- Advance the iteration index to $n + 1$ and go to the first step.

For further Reference, see Rosner et al. (2021), Section 4.4.1, and Jackman (2009).

Remark. *The Gibbs sampler is a variant of the Metropolis-Hastings algorithm in which each component of θ is updated sequentially and the acceptance probability for each step is always* 1*.*

Remark. *If we are not able to sample directly from one of the full-conditionals then it is possible to do so via the Metropolis-Hastings algorithm (Metropolis-within-Gibbs algorithm).*

Example 2.3 (see Hoff (2009), Chapter 6)**.** *We have collected n measurements of the wing length of different midges (of the same species): $y_1, \ldots, y_n$. Let us assume the following:*

$$Y_1, \ldots, Y_n | \theta, \tau \overset{iid}{\sim} \mathcal{N}\left(\theta, \frac{1}{\tau}\right)$$

The prior for the bidimensional parameter (θ, τ) is such that θ and τ are a priori independent and

$$\theta \sim \mathcal{N}\left(\mu_0, t_0^2\right), \quad \tau \sim gamma\left(\frac{\nu_0}{2}, \frac{\nu_0 s_0^2}{2}\right)$$

*To build the Gibbs sampler, we need to derive the **full-conditionals**: the conditional law of θ, given τ and data $\boldsymbol{y}$, $[\theta|\tau, \boldsymbol{y}]$ and the conditional law of τ, given θ and data $\boldsymbol{y}$, $[\tau|\theta, \boldsymbol{y}]$. In particular, we observe that both distributions are proportional to the joint law, up to a multiplicative constant:*

$$\mathcal{L}(\theta, \tau, \boldsymbol{Y}) = \mathcal{L}(\boldsymbol{y}; \theta, \tau)\, \pi(\theta)\, \pi(\tau) = \tau^{\frac{n}{2}} \mathrm{e}^{-\frac{\tau}{2}\sum_{i=1}^n (y_i-\theta)^2} \times \mathrm{e}^{-\frac{(\theta-\mu_0)^2}{2t_0^2}} \times \tau^{\frac{\nu_0}{2}-1} \mathrm{e}^{-\tau\frac{\nu_0 s_0^2}{2}} \mathbb{I}_{(0,+\infty)}(\tau)$$

Therefore, we have that:

- $[\theta|\tau, \boldsymbol{y}] \propto \mathrm{e}^{-\frac{\tau}{2}\sum_{i=1}^n (y_i-\theta)^2} \mathrm{e}^{-\frac{(\theta-\mu_0)^2}{2t_0^2}} = \mathrm{e}^{-\frac{1}{2}\left(n\tau+\frac{1}{t_0^2}\right)(\theta-\theta_n)^2}$ *with* $\theta_n = \frac{n\tau\overline{y}+\frac{\mu_0}{t_0^2}}{n\tau+\frac{1}{t_0^2}}$ *so that*

$$\theta|\tau, \boldsymbol{y} \sim \mathcal{N}\left(\theta_n, \left(n\tau + \frac{1}{t_0^2}\right)^{-1}\right).$$

- $[\tau|\theta, \boldsymbol{y}] \propto \tau^{\frac{\nu_0+n}{2}-1} \mathrm{e}^{-\frac{\tau}{2}\left[\sum_{i=1}^n (y_i-\theta)^2 + \nu_0 s_0^2\right]} \mathbb{I}_{(0,+\infty)}(\tau)$ *so that*

$$\tau|\theta, \boldsymbol{y} \sim gamma\left(\frac{\nu_0+n}{2}, \frac{1}{2}\left[\sum_{i=1}^n (y_i-\theta)^2 + \nu_0 s_0^2\right]\right).$$

It is now possible to implement the Gibbs sampler as shown in the R notebook.

2.6 Convergence Diagnostics

Most methods for assessing convergence rely on informal evaluations of individual parameters or functions of parameters. With a small number of parameters of interest, monitoring each individually is not too burdensome. On the other hand, if there are many parameters and one does not monitor all of them, one may be fooled into thinking that the process has converged based on the apparent convergence of the subset under examination.

Remark. *Many convergence diagnostics are available in the R package* `coda`*.*

Trace Plots

After we carry out an iterative method to generate samples from the posterior distribution, we will be interested in showing that the method did, in fact, converge. If the algorithm has not converged, then the generated values may not be random samples from the posterior distribution. The easiest way involves plotting the history of the generated iterates for each parameter or variable, starting with the first saved value. We produce these trace plots by graphing the iterates as a time series, with iteration number on the x-axis and the iterate itself on the y-axis, to show the history of the

consecutively generated random samples. The trajectories should eventually look like white noise without a discernible trend. Moreover, starting chains at different initial values is a way to show that we have reached convergence. That is, we should run two or more parallel chains, each starting from a different initial value. If the algorithm converges then, regardless of the initial values, all of the chains will generate random samples from the same target distribution. If plotted on the same graph, the individual trace plots should merge together at some point and then stay together. If the chains remain separate after several thousand iterations, then there is potential for a convergence problem.

Markov Chain Standard Error

An important diagnostics tool for an MCMC method is the computation of the Markov chain standard error, which is given by the standard deviation of the *estimator* $\bar{h}_T$ of the target $\bar{h} := \int h(\theta) d\pi(\theta|\boldsymbol{y})$, where $h : \Theta \to \mathbb{R}$,

$$\bar{h}_T = \frac{1}{T} \sum_{j=1}^{T} h(\theta^{(j)}), \tag{8}$$

where $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\}$ is the MC simulated via the MCMC algorithm. We can assume that we have got rid of the first BI iterations as the burn-in, and hence that we have *shifted* the time so that, marginally, each $\theta^{(j)}$ has (approximately) distribution $\pi(\theta|\boldsymbol{y})$.

If we go back to the Central Limit Theorem for ergodic MCs (see Theorem 2.9, applied to case where π is the posterior distribution of the vector of unknown parameters in the Bayesian model), the variance of the error $\bar{h}_T - \bar{h}$ (equal to the variance of the estimator $\bar{h}_T$) is expressed by

$$\frac{\sigma_h^2}{T} = \frac{1}{T} Var_\pi \left(h\left(X_0\right)\right) + \frac{2}{T} \sum_{k=1}^{+\infty} \mathrm{Cov}_\pi \left(h\left(X_0\right), h\left(X_k\right)\right)$$

in (6). We now focus on understanding how this expression has come out, and how can we indeed calculate this variance in our MCMC algorithms.

First of all, let us compute now T times the variance of (8):

$$\sigma_h^2 := T \,\mathrm{Var}\left(\bar{h}_T\right) = T \,\mathrm{Var}\left(\frac{1}{T} \sum_{j=1}^{T} h(\theta^{(j)}) \right) = T \frac{1}{T^2} \,\mathrm{Var}\left(\sum_{j=1}^{T} h(\theta^{(j)}) \right)$$

$$= \frac{1}{T} \left(\mathrm{Var}(h(\theta^{(1)})) + \mathrm{Var}(h(\theta^{(2)})) + \cdots + \mathrm{Var}(h(\theta^{(T)})) + 2 \sum_{i<j} \mathrm{Cov}(h(\theta^{(i)}), h(\theta^{(j)})) \right)$$

Now, because the marginal distributions of the $\theta^{(j)}$'s are the same, then we have

$$\sigma_h^2 := T \,\mathrm{Var}\left(\bar{h}_T\right) = \frac{1}{T} \left(T \,\mathrm{Var}(h(\theta^{(1)})) + \sum_{j=1}^{T-1} (T-j) \,\mathrm{Cov}(h(\theta^{(1)}), h(\theta^{(j+1)})) \right).$$

By definition of the correlation between two random variables, we have that

$$\mathrm{Cov}(h(\theta^{(1)}), h(\theta^{(j+1)})) = \sigma^2 \times \rho_j, \;\; j = 1, 2, 3, \ldots$$

where ρ_j is the correlation between $h(\theta^{(1)})$ and $h(\theta^{(j+1)})$ and σ^2 is the variance of each variable $h(\theta^{(j)})$. Summing up:

$$\sigma_h^2 = \sigma^2 + 2 \sum_{j=1}^{T-1} \frac{T-j}{T} \sigma^2 \rho_j \simeq \sigma^2 \left(1 + 2 \sum_{j=1}^{T} \rho_j \right) \simeq \sigma^2 \left(1 + 2 \sum_{j=1}^{+\infty} \rho_j \right)$$

and the variance of the estimator $\bar{h}_T$ is

$$\frac{\sigma_h^2}{T} \simeq \frac{\sigma^2}{T} (1 + 2 \sum_{j=1}^{+\infty} \rho_j).$$

Note that

$$\frac{\sigma_h^2}{T} \geq \frac{\sigma^2}{T} \text{ if and only if } \sum_{j=1}^{+\infty} \rho_j \geq 0$$

which typically occurs in the MCMC algorithms. Note also that $\frac{\sigma^2}{T}$ is the variance of the simple Monte Carlo estimator.
An estimate of the variance of the MCMC estimator (and hence and estimate of the MCMC standard error) is

$$\widehat{\left(\frac{\sigma_h^2}{T}\right)} = \frac{\sigma^2}{T}\widehat{(1 + 2\sum_{j=1}^{+\infty} \rho_j)} = \frac{\hat{\sigma}^2}{T}\left(1 + 2\sum_{j=1}^{M} \hat{\rho}_j\right),$$

where a large enough M is typically computed by R (R package `coda`). As an estimate $\hat{\rho}_j$ of the correlation between $h(\theta^{(1)})$ and $h(\theta^{(j+1)})$, we (and R) consider the ratio between the empirical covariances and the empirical variance

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0},$$

where

$$\hat{\gamma}_j = \frac{1}{T}\sum_{i=1}^{T-j}\left(h(\theta^{(i)}) - \bar{h}_T\right)\left(h(\theta^{(i+j)}) - \bar{h}_T\right),\ j = 0, 1, 2, \ldots$$

Note that $\hat{\gamma}_0$ is the empirical variance of the $h(\theta^{(i)})$'s.
Good values for $\widehat{\left(\frac{\sigma_h^2}{T}\right)}$ (which can be computed through the R command `batchSE`) should be small. In particular, the ratio between MCse (of a target) and the posterior standard deviation has to be typically less than 0.1%, 1% or 5%.

Another important diagnostic to compute is the **effective sample size** (command `effectiveSize`), which is defined as the number of independent samples we should simulate **iid** from the posterior distribution to get the same Markov chain standard error that we have actually obtained. More precisely, the effective sample size is the (positive) value $\hat{n}$ such that

$$\widehat{\text{Var}(\bar{h}_T)} = \frac{\hat{\sigma}^2}{\hat{n}},$$

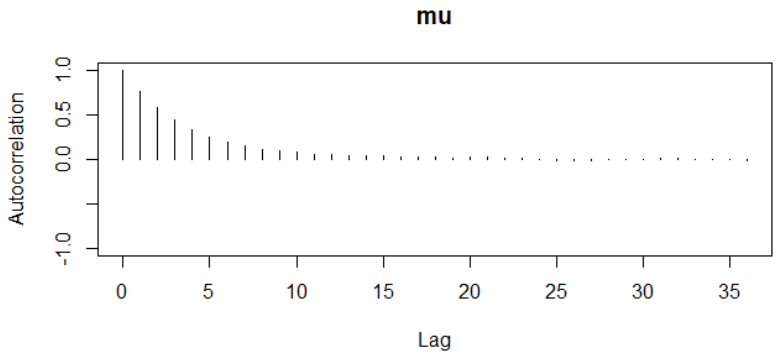
In particular, the higher is this value (the closer to the number of simulated MCMC draws), the better, as it represents the effective number of independent samples that have been simulated.

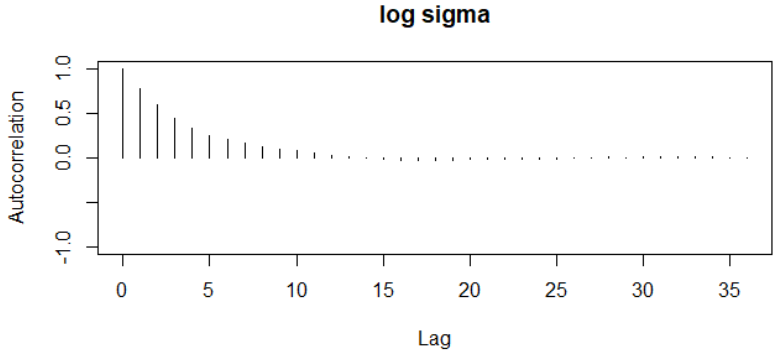
Remark. *Both the Markov chain standard error and the effective sample size are specific of the unidimensional components of the parameter vector.*

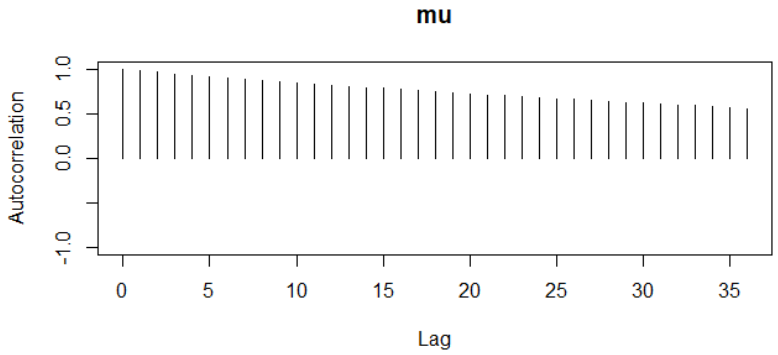
Autocorrelation Plots

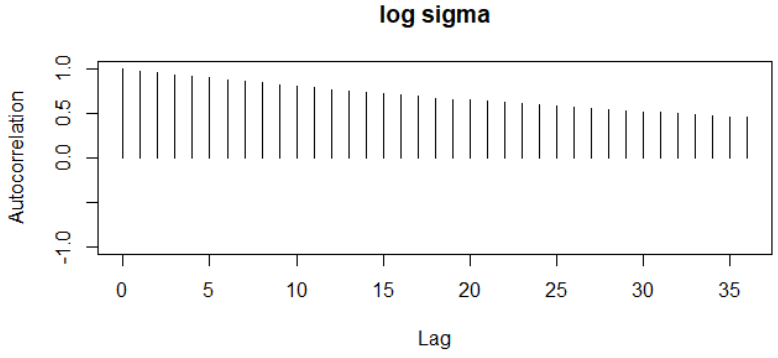
As a further diagnostic for the *good mixing* of the chain, we typically focus on the *autocorrelation plots*, i.e. the barplots of $\hat{\rho}_0 = 1, \hat{\rho}_1, \hat{\rho}_2, \ldots$.
We expect the autocorrelation to decrease as the lag increases. In other words, if the chain is mixing well, then there should be (much) less correlation between values 20 iterations apart than between values only one iteration apart. The following autocorrelation plots address both the case in which the chain is well-behaved and the case in which is not well-behaved.
A problem caused by autocorrelation in the sample is that the effective sample size is generally smaller, and sometimes much smaller, than the number of samples that are kept.

For further details on convergence diagnostics, see Rosner et al. (2021), Section 4.4.5.


mu
Autocorrelation
1.0
0.5
0.0
-1.0
0
5
10
15
20
25
30
35
Lag


log sigma
Autocorrelation
1.0
0.5
0.0
-1.0
0
5
10
15
20
25
30
35
Lag


mu
Autocorrelation
1.0
0.5
0.0
-1.0
0
5
10
15
20
25
30
35
Lag


log sigma
Autocorrelation
1.0
0.5
0.0
-1.0
0
5
10
15
20
25
30
35
Lag

3 Bayesian Linear Models

3.1 The likelihood in the Linear Regression Model

Regression models are defined mathematically with a functional relationship between the mean of a response variable Y and some predictor variables $\boldsymbol{x}$ (also called **covariates**). Let us assume that the data consist of a set of observations for n statistical units: $\{(y_i, x_{i1}, \ldots, x_{ik}), i = 1, \ldots, n\}$. We assume that

$$Y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2), i = 1, \ldots, n.$$

To keep the model simple, we have assumed that the conditional variance of the response Y_i's does not vary with the subject i. The model can be equivalently described as

$$\boldsymbol{Y}|X, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}_n\left(\mathbb{E}\left[\boldsymbol{Y}|X, \boldsymbol{\beta}\right], \sigma^2\mathbb{I}_n\right) \tag{9}$$

$$\mathbb{E}\left[\boldsymbol{Y}|X, \boldsymbol{\beta}\right] = X\boldsymbol{\beta} \tag{10}$$

where $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$ and $X = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$.

The parameters of interest are $(\boldsymbol{\beta}, \sigma^2)$ or $(\boldsymbol{\beta}, \tau)$ where $\tau := \dfrac{1}{\sigma^2}$ is the precision at the observation level.

Remark. *Although the predictor variables often result from random sampling in practice, we will treat them as fixed, denoting them using lower-case letters* $(x_{i1}, \ldots, x_{ik})$*. Note that, if they are random, but we assume that, a priori,* X *and* $(\boldsymbol{\beta}, \sigma^2)$ *are independent, then the conditional distribution of* X *and data* $\boldsymbol{Y}$ *can be expressed as:*

$$\begin{aligned} f(\boldsymbol{y}, X|\omega, \boldsymbol{\beta}, \sigma^2) &= f(\boldsymbol{y}|X, \omega, \boldsymbol{\beta}, \sigma^2)p(X|\omega, \boldsymbol{\beta}, \sigma^2) \\ &= f(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2)p(X|\omega); \end{aligned}$$

hence, since the interest here is in $(\boldsymbol{\beta}, \sigma^2)$*,* $p(X|, \omega)$ *represents a constant factor in the likelihood (and in the posterior) and therefore it can be discarded.*

The likelihood can be written as

$$L\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \mathrm{e}^{-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}}$$

if the parameters of interest are $(\boldsymbol{\beta}, \sigma^2)$ or

$$L\left(\boldsymbol{\beta}, \tau; \boldsymbol{y}\right) \propto \tau^{\frac{n}{2}} \mathrm{e}^{-\frac{\tau}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}$$

for the parameters $(\boldsymbol{\beta}, \tau)$. Let us recall what are the maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ and σ^2. Assuming that $p := k+1 < n$ and that the matrix X is full rank (so that X^TX is invertible) we have that:

- The maximum likelihood estimate of $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}} := \left(X^TX\right)^{-1}X^T\boldsymbol{y}$ and its covariance is given by $\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}\middle|\sigma^2\right) = \sigma^2\left(X^TX\right)^{-1}$.
- An unbiased estimator for σ^2 is given by $\widehat{\sigma}^2 := \frac{1}{n-p}\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right) = \frac{s^2}{n-p}$ where $s^2 = \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)$ is the sum of squared residuals.
- Confidence interval for each regression parameter β_i are built through the pivotal quantity $T_i := \frac{\widehat{\beta}_i - \widehat{\beta}}{\sqrt{\sigma^2\omega_{ii}}} \sim t\left(n-p\right)$ where we have set $\left(X^TX\right)^{-1} = [\omega_{ij}]$.

Having introduced such quantities it is possible to re-write the likelihood as follows:

$$\begin{aligned}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta}) &= \left(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\right)+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T X^TX\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\\ &= s^2+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T X^TX\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\end{aligned}$$

so that

$$L\left(\boldsymbol{\beta},\sigma^2;\boldsymbol{y}\right)\propto\frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}}\mathrm{e}^{-\frac{s}{2\sigma^2}-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^TX^TX\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}$$

or

$$L\left(\boldsymbol{\beta},\tau;\boldsymbol{y}\right)\propto\tau^{\frac{n}{2}}\mathrm{e}^{-\frac{\tau s^2}{2}-\frac{\tau}{2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^TX^TX\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}$$

3.2 Priors and posteriors

We now discuss different types of prior specifications of $(\boldsymbol{\beta},\tau)$ in the linear regression model.

Conjugate Prior for $\boldsymbol{\beta}$ when σ^2 (or τ) is known
If σ^2 is known the conjugate prior for $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta}\sim\mathcal{N}\left(\boldsymbol{b}_0,\mathbb{B}_0\right)$ where $\mathbb{B}_0$ is an invertible $p\times p$ matrix. Indeed, by using such prior we get:

$$\pi\left(\boldsymbol{\beta}|\boldsymbol{y}\right)\propto\mathrm{e}^{-\frac{\tau}{2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^TX^TX\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{b}_0\right)^T\mathbb{B}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{b}_0\right)}$$

which implies that the posterior is

$$\boldsymbol{\beta}|\boldsymbol{y}\sim\mathcal{N}\left(\left(\tau X^TX+\mathbb{B}_0^{-1}\right)^{-1}\left(\tau X^TX\widehat{\boldsymbol{\beta}}+\mathbb{B}_0^{-1}\boldsymbol{b}_0\right),\left(\tau X^TX+\mathbb{B}_0^{-1}\right)^{-1}\right)$$

In particular, we have that the posterior mean is given by a weighted average of the prior mean $\boldsymbol{b}_0$ and the frequentist estimate $\widehat{\boldsymbol{\beta}}$. Moreover, observe that it is not necessary to assume that X^TX is invertible as it suffices to have $\mathbb{B}_0$ invertible.

Conjugate Prior for $\boldsymbol{\beta}$ and σ^2: Normal-invgamma Prior
Observe that in the expression $L\left(\boldsymbol{\beta},\sigma^2;\boldsymbol{y}\right)\propto\frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}}\mathrm{e}^{-\frac{s}{2\sigma^2}-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^TX^TX\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}$ the parameter $\boldsymbol{\beta}$ occurs as in the kernel of a Gaussian distribution whereas the parameter $\frac{1}{\sigma^2}$ occurs as in the kernel of an inverse-gamma distribution. Therefore the conjugate prior is given by

$$\pi\left(\boldsymbol{\beta},\sigma^2\right)=\pi\left(\boldsymbol{\beta}|\sigma^2\right)\pi\left(\sigma^2\right)$$

where

$$\boldsymbol{\beta}|\sigma^2\sim\mathcal{N}\left(\boldsymbol{b}_0,\sigma^2\mathbb{B}_0\right),\quad\sigma^2\sim\text{inv-gamma}\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)\tag{11}$$

with $\mathbb{B}_0$ invertible $p\times p$ matrix and $\nu_0,\sigma_0^2>0$. In this case the posterior is given by

$$\pi\left(\boldsymbol{\beta},\sigma^2|\boldsymbol{y}\right)=\pi\left(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}\right)\pi\left(\sigma^2|\boldsymbol{y}\right)$$

where

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},X\sim\mathcal{N}\left(\boldsymbol{b}_n,\sigma^2\mathbb{B}_n\right),\quad\sigma^2|\boldsymbol{y},X\sim\text{inv-gamma}\left(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2}\right)$$

with

$$\begin{aligned}
\mathbb{B}_n &= \left(X^T X + \mathbb{B}_0^{-1}\right)^{-1} \\
\boldsymbol{b}_n &= \mathbb{B}_n \left(X^T X \widehat{\boldsymbol{\beta}} + \mathbb{B}_0^{-1} \boldsymbol{b}_0\right) \\
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \frac{1}{\nu_n} \left(\nu_0 \sigma_0^2 + s^2 + \left(\boldsymbol{b}_0 - \widehat{\boldsymbol{\beta}}\right)^T \left(\mathbb{B}_0 + \left(X^T X\right)^{-1}\right)^{-1} \left(\boldsymbol{b}_0 - \widehat{\boldsymbol{\beta}}\right)\right)
\end{aligned}$$

Moreover, the marginal posterior for $\boldsymbol{\beta}$, obtained integrating out the posterior of σ^2, is given by

$$\boldsymbol{\beta}|\boldsymbol{y}, X \sim t_p\left(\boldsymbol{b}_n, \sigma_n^2 \mathbb{B}_n, \nu_n\right)$$

(multivariate t distribution with location given by $\boldsymbol{b}_n$, scale given by $\sigma_n^2 \mathbb{B}_n$ and ν_n degrees of freedom).

If we want to obtain predictions for *new* subjects joining the sample, let be $\boldsymbol{Y}_{\text{new}}$ the vector of m new observations and the corresponding new $m \times p$ covariate matrix X_{new}, the predictive distribution for $\boldsymbol{Y}_{\text{new}}$ is given by

$$\boldsymbol{Y}_{\text{new}}|\boldsymbol{y}, X, X_{\text{new}} \sim t_m\left(X_{\text{new}} \boldsymbol{b}_n, \sigma_n^2 \left(\mathbb{I}_m + X_{\text{new}} \mathbb{B}_n X_{\text{new}}^T\right), \nu_n\right).$$

Here $\mathbb{I}_m$ is the m-dimensional identity matrix.

Zellner's g-Prior
The most difficult aspect of the choice of the hyperparameters of the Normal-Gamma prior is the derivation of the matrix $\mathbb{B}_0$. Zellner proposed the following prior:

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}\left(\boldsymbol{b}_0, \sigma^2 \mathbb{B}_0\right), \mathbb{B}_0 = c\left(X^T X\right)^{-1}, \\
\sigma^2 \sim \text{inv-gamma}\left(\frac{0}{2}, \frac{0 \times \sigma_0^2}{2}\right) \propto \frac{1}{\sigma^2} \mathbb{I}_{(0,+\infty)}\left(\sigma^2\right) \text{ OR } \tau \propto \frac{1}{\tau} \mathbb{I}_{(0,+\infty)}(\tau)
\end{aligned}$$

Here c **is a positive constant**.

Remark. *Zellner's prior corresponds to the conjugate prior with $\mathbb{B}_0 = c\left(X^T X\right)^{-1}$ and $\nu_0 = 0$; it is an improper prior (we do not like that ν_0 is assumed to be 0). Moreover, it requires $X^T X$ to be invertible.*

By applying the general formula for the posterior introduced in the previous chapter we get

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, X &\sim \mathcal{N}\left(\boldsymbol{b}_n, \sigma^2 \mathbb{B}_n\right) \\
\sigma^2|\boldsymbol{y}, X &\sim \text{inv-gamma}\left(\frac{n}{2}, \frac{1}{2}\left(s^2 + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{b}_0\right)^T \mathbb{B}_n^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{b}_0\right)\right)\right)
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{B}_n &= \frac{c}{c+1}\left(X^T X\right)^{-1} \\
\boldsymbol{b}_n &= \frac{1}{c+1} \boldsymbol{b}_0 + \frac{c}{c+1} \widehat{\boldsymbol{\beta}}
\end{aligned}$$

Note that the posterior mean is a weighted average of the prior mean $\boldsymbol{b}_0$ and the MLE $\widehat{\boldsymbol{\beta}}$. In particular, as the parameter c increases, the weight given to the observed data (represented by $\widehat{\boldsymbol{\beta}}$) $\boldsymbol{b}_n$ increases. Hence, for instance, $c = 1$ corresponds to equal weights. We can select the value of c according to this criterion. However, often c is fixed as $c = \log(n)$.

This prior is known as Zellner's g-prior since the constant c was denoted as g in the paper by Zellner where it was proposed first.

Reference Prior
We introduce here the flat prior for $(\boldsymbol{\beta}, \sigma^2)$

$$\pi\left(\boldsymbol{\beta}, \sigma^2\right) \propto \frac{1}{\sigma^2}\mathbb{I}_{(0,+\infty)}\left(\sigma^2\right)$$

which is equivalent to

$$\pi\left(\boldsymbol{\beta}, \tau\right) \propto \frac{1}{\tau}\mathbb{I}_{(0,+\infty)}\left(\tau\right)$$

for parameterization $(\boldsymbol{\beta}, \tau)$. We consider it in this Lecture Notes only because it leads to frequentist estimates. However, students are encouraged to adopt proper priors, possibly informative.

In this case, it is straightforward to compute the posterior:

$$(\boldsymbol{\beta}, \tau|\boldsymbol{y}) = \pi\left(\boldsymbol{\beta}|\tau, \boldsymbol{y}\right)\pi\left(\tau|\boldsymbol{y}\right)$$

where $\boldsymbol{\beta}|\tau, \boldsymbol{y}, X \sim \mathcal{N}\left(\widehat{\boldsymbol{\beta}}, \frac{1}{\tau}\left(X^TX\right)^{-1}\right)$ and $\tau|\boldsymbol{y}, X \sim \text{gamma}\left(\frac{n-p}{2}, \frac{s^2}{2}\right)$.

Remark. *Observe that the posterior is proper if $n > p$ and X^TX is invertible.*

In this case, if $\boldsymbol{Y}_{\text{new}}$ is the vector of m new observations and X_{new} is the corresponding new $m \times p$ covariate matrix, the predictive distribution for $\boldsymbol{Y}_{\text{new}}$ is given by

$$\boldsymbol{Y}_{\text{new}}|\boldsymbol{y}, X, X_{\text{new}} \sim t_m\left(X_{\text{new}}\widehat{\boldsymbol{\beta}}, \frac{s^2}{n-p}\left(\mathbb{I}_m + X_{\text{new}}\left(X^TX\right)^{-1}X_{\text{new}}^T\right), n-p\right).$$

Conditionally Conjugate Independence Prior
The most common choice for the prior of $(\boldsymbol{\beta}, \tau)$ is to assume that $\boldsymbol{\beta}$ and τ are a priori independent and that

$$\boldsymbol{\beta} \sim \mathcal{N}_p\left(\boldsymbol{b}_0, \mathbb{B}_0\right), \ \tau \sim \text{gamma}\left(a, b\right)$$

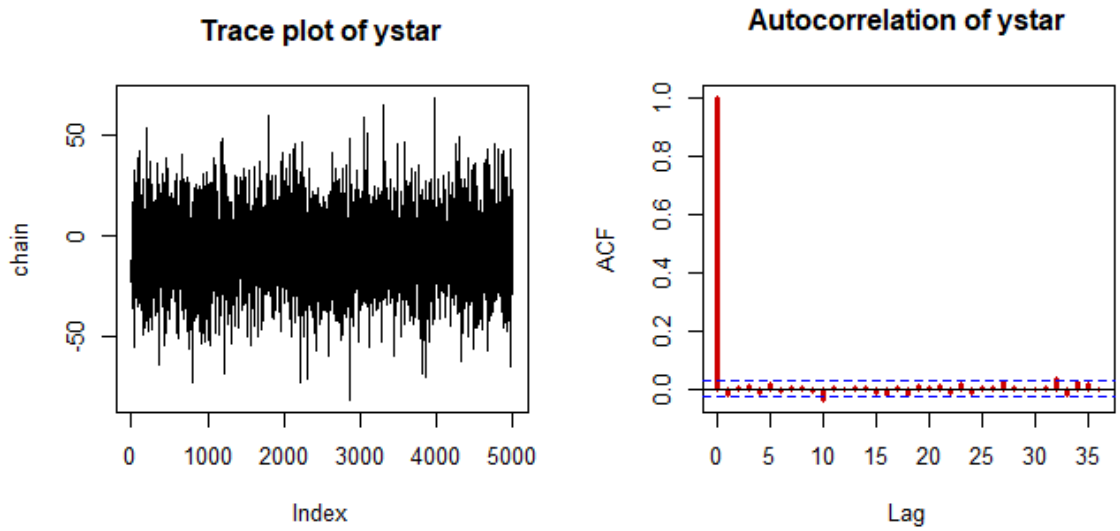
$\mathbb{B}_0$ is an invertible $p \times p$ matrix. This prior is **semi-conjugate**, i.e. the two full-conditionals of the Gibbs sampler are in closed form. Specifically, the full conditional of $\boldsymbol{\beta}$ is a p-dimensional Gaussian, and the full-conditional of τ is a gamma density.

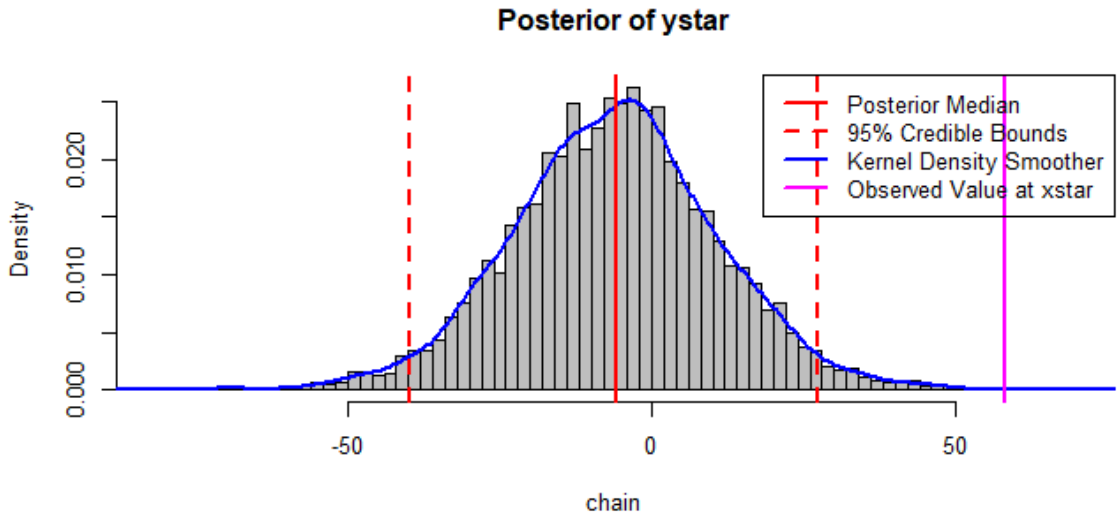
Example 3.1 (see Jackman (2009), Example 2.15)**.** *In November 1993, the state of Pennsylvania conducted elections for its state legislature. The result in the Senate election in the 2nd district (based in Philadelphia) was challenged in court, and ultimately overturned. The Democratic candidate won* $19,127$ *of the votes cast by voting machine, while the Republican won* $19,691$ *votes cast by voting machine, giving the Republican a lead of* 564 *votes. However, the Republican won just* 371 *absentee ballots whereas the Democrat won* $1,396$ *absentee ballots, enough offset the Republican lead based on the votes recorded by machines on election day. Therefore, the Republican candidate sued, claiming that many of the absentee ballots were fraudulent, and the judge in the case solicited expert analysis from Orley Ashenfelter (an economist at Princeton University). Ashenfelter examined the relationship between absentee vote margins and machine vote margins in* 21 *previous Pennsylvania Senate elections in seven districts in the Philadelphia area over the preceding decade. This is part of his analysis, obtained by R and JAGS (Just Another Gibbs Sampler). See* `Jackman_exampleLinearRegression_JAGS.Rmd` *and the text file* `regression.bug` *describing the Bayesian model.*

```
library(rjags)
## Creation of the variables for the regression analysis
y = (absdem - absrep)/(absdem + absrep)*100 # Democratic Margin, Absentee
    Ballots (Percentage Points)
x = (machdem - machrep)/(machdem + machrep)*100 # Democratic Margin,
   Machine Ballots (Percentage Points)
x.susp = x[22]
y.susp = y[22]
y = y[1:21]
```

```
x = x[1:21]
## Define the data: output vector (y), covariates vector (x), sample size
    (n) and suspect covariate value (xstar)
data = list(y=y[1:21],x=x[1:21],n=21,xstar=x.susp)
## Define the list of initial values for the MCMC algorithm that JAGS
    will implement
inits = function() {list(beta=c(0,0),sigma=5,ystar=0)}
## The Bayesian model is in the text file "regression.bug". This
## file, in addition to the list "data", is taken as an input
## by JAGS for generating the MCMC in 3 steps:
## - The first step (jags.model) gets all the info into JAGS and
## let JAGS figure out appropriate sampler for the model.
## - The second step (update) runs the chain for a burn-in period.
## - The third step (coda.samples) runs and records the MCMC sample
## we will subsequently examine.
modelRegress = jags.model("regression.bug",data=data,inits=inits,n.adapt
    =1000,n.chains=1)
## File "regression.bug" specifies both the likelihood and the prior.
## In particular, beta has 2 components which are iid Gaussian with mean
    = 0 and
## variance = 10,000 whereas sigma is Uniform on the interval (0,100)
## Observe that the prior is not conjugate.
update(modelRegress,n.iter=19000) # This is the burn-in period and it is
    not stored
variable.names = c("beta","sigma","ystar")
n.iter = 50000
thin = 10 # We decide to store just 10% of the generated samples
outputRegress = coda.samples(model=modelRegress,variable.names=variable.
    names,n.iter=n.iter,thin=thin)
## Posterior Predictive Distribution
data.out = as.matrix(outputRegress)
data.out = data.frame(data.out)
ystar.pred = data.out[,'ystar']
x11()
chain = ystar.pred
layout(matrix(c(1,2,3,3),2,2,byrow=T))
plot(chain,type="l",main="Trace plot of ystar")
acf(chain,lwd=3,col="red3",main="Autocorrelation of ystar")
hist(chain,nclass="fd",freq=F,main="Posterior of ystar",col="gray")
lines(density(chain),col="blue",lwd=2,xlim=c(0,60)) # Kernel Density
    Estimate
quantile(chain,prob=c(0.025,0.5,0.975)) # Posterior Credible Interval
abline(v=quantile(chain,prob=c(0.025)),col="red",lty=2,lwd=2)
abline(v=quantile(chain,prob=c(0.5)),col="red",lty=1,lwd=2)
abline(v=quantile(chain,prob=c(0.975)),col="red",lty=2,lwd=2)
abline(v=y.susp,col="magenta",lwd=2)
legend("topright",legend=c("Posterior Median", "95% Credible Bounds","
    Kernel Density Smoother","Observed Value at xstar"),lwd=c(2,2,2,2),
    col=c("red","red","blue","magenta"),lty=c(1,2,1,1))
```

The result of this script is shown in the following plot.





In particular, we have that the value of the Democratic margin observed in the case of the elections in analysis is pretty far from the region in which the posterior probability density concentrates its mass. The corresponding JAGS code is given below.

```
model{
for(i in 1:n) {
        mu[i] <- beta[1] + beta[2]*x[i] #<- is a deterministic
            assignment
        y[i] ~ dnorm(mu[i],tau) # Mind thattau is the PRECISION of the
            Gaussian distribution
    }

    ## Priors
    beta[1] ~ dnorm(0,.0001)
    beta[2] ~ dnorm(0,.0001)
    sigma ~ dunif(0,100)
    tau <- pow(sigma,-2)
```

```
    ## Suspect case
    mustar <- beta[1] + beta[2]*xstar
    ystar ~ dnorm(mustar, tau)
}
```

3.3 Generalized Linear Models

Generalized Linear Models (GLMs) are a flexible generalization of ordinary linear regression. In this case each outcome Y of the dependent variable is assumed to be generated from a particular distribution in an exponential family (a large class of probability distributions that includes the normal, binomial, Poisson and gamma distributions) and the mean μ of the distribution depends on the independent variables $\boldsymbol{x}$ through $\mathbb{E}[Y|\boldsymbol{x}] = \mu = h(x\boldsymbol{\beta})$ where h is some real-valued function. A GLM consists of the following three elements:

- The **random component**, which is given by the distribution of Y_i, given $\boldsymbol{x}_i$, within an exponential family of probability distributions: $Y_i|\boldsymbol{x}_i \sim Exp - family$,

  $$f(y_i|\theta_i, \varphi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right\}$$

  where θ_i is the **natural parameter** and φ is the **scale parameter**. We demote by μ_i the expectation $\mathbb{E}[Y_i|\boldsymbol{x}_i]$

- The **linear predictor** $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, i.e., we choose the covariates.

- The **link function** g such that

  $$g(\mu_i) = \eta_i, \text{ that is } g(\mathbb{E}[Y_i|\boldsymbol{x}_i]) = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

  Equivalently, the link between the expectation of Y_i and the linear predictor η_i can be expressed by the **response function** h, that is the inverse of the link function g:

  $$\mu_i = h(\eta_i) = h(x_i^T\boldsymbol{\beta}).$$

Remark. *The natural link function is the link function we get equating the natural parameter to the liner predictor.*

Note that parameter θ_i will depend on the linear predictor $\eta_i = x_i^T\boldsymbol{\beta}$. The scale parameter φ may be present in the random component or not. All in all, the parameters of interest in the case of GLMs are given by $\boldsymbol{\beta}$ and φ and the most common priors are the followings:

- $\pi(\boldsymbol{\beta}, \varphi) = \pi(\boldsymbol{\beta}|\varphi)\pi(\varphi)$.
- $\pi(\boldsymbol{\beta}, \varphi) = \pi(\boldsymbol{\beta})\pi(\varphi)$.

The typical prior for $\boldsymbol{\beta}$ is $\mathcal{N}_p(\boldsymbol{b}_0, B_0)$, where B_0 is a $p \times p$ matrix. The prior mean is often chosen as $\mathbf{0}$ (if the response is at least centered and B_0 is the covariance matrix given by a constant c (larger than 1, e.g., 10) times a frequentist estimate, e.g., the covariance matrix of the MLE.

3.3.1 Binary Response Regression

In binary response regression we have $Y \in \{0, 1\}$ and the random component is given by the Bernoulli distribution: $Y_i|\boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Be}(\pi_i) = \text{Be}(\pi(\boldsymbol{x}_i))$. Therefore, we have that $\mu_i = \pi_i \in (0, 1)$ for all i. Consequently, a proper choice for the link function is given by $\pi_i = F\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$ where F is a cumulative distribution function. The most common choices for F are the followings:

- In the **probit model** we have $F(x) = \Phi(x)$ (Φ is the cumulative distribution function of the standard normal distribution).

- In the **logit model** we have $F(x) = \frac{e^x}{1+e^x}$ (F is the cumulative distribution function of the logistic distribution).
- In the **complementary log-log model** we have $F(x) = 1 - e^{-e^x}$.

Remark. *Observe that*

$$\mathbb{P}(Y_i = y|x_i) = \pi_i^y (1-\pi_i)^{1-y} = e^{y \log \pi_i + (1-y)\log(1-\pi_i)} = e^{y \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)}, \ y \in \{0,1\}$$

where π_i depends on $\boldsymbol{x}_i$. Therefore the Bernoulli distribution belongs to the exponential family of distributions with no scale parameter.

Gibbs Sampler for the Probit Model

In order to implement a Gibbs sampler in the case of the probit model it is useful to introduce the following latent variables:

$$Z_1, \ldots, Z_n \overset{\text{ind}}{\sim} \mathcal{N}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}, 1\right)$$

so that

$$Y_i = \begin{cases} 1 & Z_i > 0 \\ 0 & Z_i \leq 0 \end{cases}.$$

It is easy to show that such scenario is equivalent to the one where the Y_i's are Bernoulli distributed with parameter $\pi_i = \Phi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$. In fact, we have

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i > 0) = 1 - \mathbb{P}(Z_i < 0) = 1 - \Phi\left(-\boldsymbol{x}_i^T \boldsymbol{\beta}\right) = \Phi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$$

Hence, the parameters are $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ and $\boldsymbol{\beta}$. In order to design a Gibbs sample to sample from the posterior, we write the joint law of $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{\beta}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\beta}) &= \mathcal{L}(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{\beta}) \mathcal{L}(\boldsymbol{Z}|\boldsymbol{\beta}) \mathcal{L}(\boldsymbol{\beta}) = \\ &= \prod_{i=1}^n \left(\mathbb{I}_{(Y_i=1)}\mathbb{I}_{(Z_i>0)} + \mathbb{I}_{(Y_i=0)}\mathbb{I}_{(Z_i \leq 0)}\right) \prod_{i=1}^n \mathcal{N}\left(Z_i; \boldsymbol{x}_i^T \boldsymbol{\beta}\right) \pi(\boldsymbol{\beta}) \end{aligned}$$

Therefore, we have that the full-conditional for $\boldsymbol{Z}$ is given by:

$$[\boldsymbol{Z}|\boldsymbol{\beta}, \boldsymbol{Y}] \propto \prod_{i=1}^n \left\{ \left(\mathbb{I}_{(Y_i=1)}\mathbb{I}_{(Z_i>0)} + \mathbb{I}_{(Y_i=0)}\mathbb{I}_{(Z_i \leq 0)}\right) \mathcal{N}\left(Z_i; \boldsymbol{x}_i^T \boldsymbol{\beta}\right) \right\}$$

so that it is possible to sample each variable Z_i independently from the distribution

$$Z_i|\boldsymbol{\beta}, \boldsymbol{y} \ \sim \ [Z_i|\boldsymbol{\beta}, \boldsymbol{Y}] = \begin{cases} \mathcal{N}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}, 1\right) \mathbb{I}_{(0,+\infty)}(Z_i) & \text{if } Y_i = 1 \\ \mathcal{N}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}, 1\right) \mathbb{I}_{(-\infty,0)}(Z_i) & \text{if } Y_i = 0 \end{cases}.$$

This is a truncated normal distribution, truncated on $(0, +\infty)$ if the associated $y_i = 1$, or truncated on $(-\infty, 0)$ if the associated $y_i = 0$.

If we assume a prior for $\boldsymbol{\beta}$ given by $\mathcal{N}(\boldsymbol{b}_0, \mathbb{B}_0)$ we have that the full-conditional for $\boldsymbol{\beta}$ is given by:

$$[\boldsymbol{\beta}|\boldsymbol{Z}, \boldsymbol{Y}] \propto \prod_{i=1}^n \mathcal{N}\left(Z_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, 1\right) \pi(\boldsymbol{\beta}) = \mathcal{N}\left(\widetilde{\boldsymbol{b}}, \widetilde{\mathbb{B}}\right) \tag{12}$$

where $\widetilde{\mathbb{B}} = \left(X^T X + \mathbb{B}_0^{-1}\right)^{-1}$ and $\widetilde{\boldsymbol{b}} = \widetilde{\mathbb{B}}\left(X^T X \widehat{\boldsymbol{\beta}} + \mathbb{B}_0^{-1} \boldsymbol{b}_0\right)$. This calculation can be understood if we see that (12) is the posterior of a linear regression model where the data are the Z_i's and the prior for β is conjugate.

Remark. *If $X \sim F$, where F is a distribution function, and $U \sim \mathcal{U}(0,1)$ then*

$$T = F^{-1}\left(\left(F(b) - F(a)\right)U + F(a)\right) \text{ has the same distribution as } X, \text{ given that } X \in (a,b).$$

The associated distribution function is

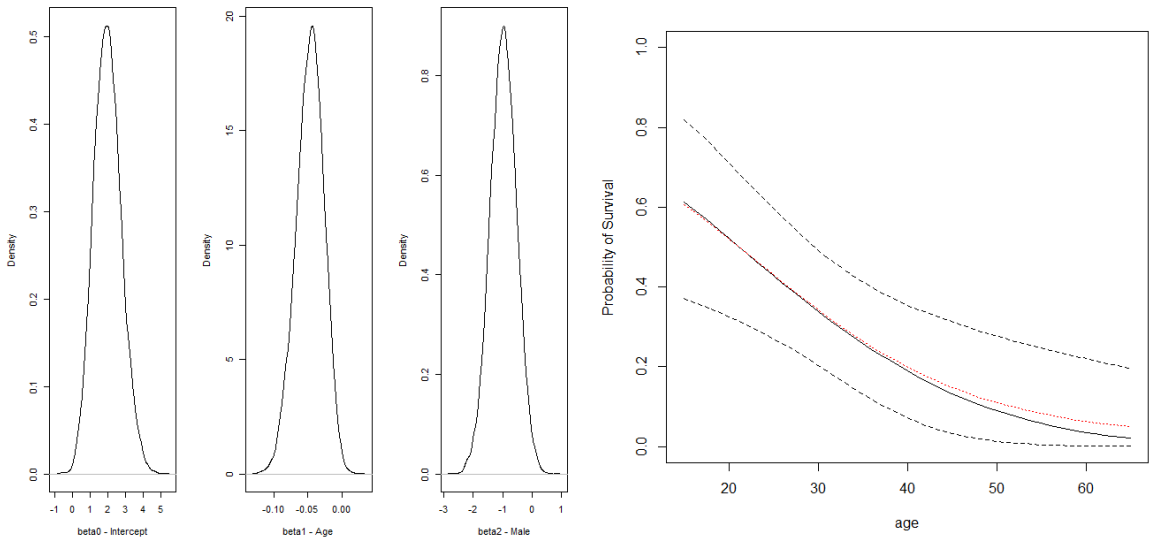
$$F_{X|\ X\in(a,b)}(t) = \begin{cases} 0 & \text{if } t < a \\ \frac{F(t)-F(a)}{F(b)-F(a)} & \text{if } a \le t < b \\ 1 & \text{if } t \ge b \end{cases}$$

See Jackman (2009), Chapter 8 and Rosner et al. (2021), Chapter 8. For Poisson regression see Rosner et al. (2021), Chapter 9.

Example 3.2 (see Albert (2009), Section 10.3)**.** *In 1846–1847 a group of $n = 45$ pioneers crossed Sierra Nevada in a wagon train and most of them starved to death. The data at our disposal consist of a binary indicator of survival y_i for each pioneer i, plus their age and gender.*

```
library(LearnBayes)
data(donner)
attach(donner)

X = cbind(1,age,male) # Definition of the design matrix. The first column
    is equal to the vector 1
fit = glm(survival~X-1,family=binomial(link=probit))
summary(fit) # Frequentist MLE estimate of the parameters in the GLM:
    1.91730001, -0.04570867, -0.95827865

# Prior for beta: N_p(b0,B0) with B0 = c0*(X^T%*%X)^-1
b0 = c(0,0,0)
c0 = 100
P0 = t(X)%*%X/c0
inv = function(X){
    EV = eigen(X)
    EV$vector%*%diag(1/EV$values)%*%t(EV$vector)
}
B0 = inv(P0)

# The Gibbs sampler is implemented in the function bayes.probit
m = 10000 # Number of simulations desired
fitBayes=bayes.probit(survival,X,m,list(beta=b0,P=P0))
# Posterior mean and standard deviation of the regression coefficients
apply(fitBayes$beta,2,mean) # 2.00342752, -0.04791917, -0.98544630
apply(fitBayes$beta,2,sd) # 0.78167631, 0.02103594, 0.45163199

x11()
par(mfrow=c(1,3))
plot(density(fitBayes$beta[,1]), xlab='beta0 - Intercept',main=' ')
plot(density(fitBayes$beta[,2]), xlab='beta1 - Age',main=' ')
plot(density(fitBayes$beta[,3]), xlab='beta2 - Male',main=' ')
mean(fitBayes$beta[,2]<0) # 0.9943
mean(fitBayes$beta[,3]<0) # 0.9889
# Both beta1 and beta2 are significative, and in both cases the
# marginal posterior is concentrated on negative values.
# Survival probability decreases as age increases and it is
# smaller for men than for women.
```

```
# Survival probability, given 'age' and 'male' values, is equal to
# Phi(beta0+beta1*age+beta2*male), i.e. it is a function of the betas.
# Let us compute its posterior distribution for a man (male=1), as age
    varies.
a = seq(15,65)
X1 = cbind(1,a,1)
p.male = bprobit.probs(X1, fitBayes$beta) # We use the function bprobit.
    probs
# We now plot for each value of age the posterior median and 90% CI.
x11()
par(mfrow=c(1,1))
plot(a,apply(p.male,2,quantile,.5),type="l",ylim=c(0,1), xlab="age",ylab=
    "Probability of Survival")
lines(a,apply(p.male,2,quantile,.05),lty=2)
lines(a,apply(p.male,2,quantile,.95),lty=2)
lines(a,apply(p.male,2,mean),lty=3,col='red') # This is the posterior
    predictive probability that Y_i^new at all ages would have survived

# Alternative way: use the MCMC pack
library(MCMCpack)
y = survival
fit = MCMCprobit(y ~ age + as.factor(male), b0=b0, B0=P0, marginal.
    likelihood="Chib95")
fitout = as.matrix(fit)
```



For a recap on linear models and details on the priors for linear models, see Rosner et al. (2021), Sections 7.1-7.4.

4 Hierarchical Models

We now consider models that have additional complexity beyond those that we have already considered. When considering regression modeling, there was a presumption that important predictor variables were taken into account. Indeed, when designing a study to assess the importance of various predictor variables on a response variable of interest, scientists will do their best to include those variables that they deem important and they are able to measure. There will be situations, however, where this has been done and where it is also suggested that there may be some other factors that:

- Were not taken into account and perhaps should have been.
- Are difficult to measure and are consequently left out of the model.
- Are perhaps even impossible to measure.

With a slight abuse of the term, we refer to such variables as **latent**. We will now focus on extending previous models to allow for latent variables. Generically, we have data y, latent variables γ and a collection of parameters θ. We pose a model for the data given the parameters and latents, $p(y|\theta,\gamma)$, a model for the latents, $p(\gamma|\theta)$, and a prior distribution, $\pi(\theta)$. This leads to a joint distribution given by:

$$\mathcal{L}(y,\gamma,\theta) = p(y|\theta,\gamma)\ \ p(\gamma|\theta)\,\pi(\theta)$$

from which it is possible to obtain the posterior probability distribution for (θ,γ):

$$p(\theta,\gamma|y) \propto \mathcal{L}(y,\gamma,\theta)$$

The model as described thus far is a generic version of a two-level hierarchical model for the data, with a third level corresponding to the prior. Such models are also termed **multilevel models**. If we have a vector z of future data with model $p(z|\theta,\gamma,y)$ (where we do not necessarily assume that Z is independent of Y conditioning on (θ,γ)) then the predictive density is defined as:

$$p(z,y) = \int p(z|\theta,\gamma,y)\,p(\theta,\gamma|y)\,\mathrm{d}\theta\mathrm{d}\gamma$$

Theoretically, it is quite simple to make a full range of statistical inferences based on this model. It is often simple in practice, as well, provided we are able to iteratively sample from the conditional distributions:

$$p(\theta|\gamma,y), p(\gamma|\theta,y), p(z|\theta,\gamma,y)$$

Example 4.1 (ANOVA model - see Hoff (2009), Section 8.4)**.** *We now consider a survey collecting the math scores of students from* 100 *different public high schools in the US. Such data can be analyzed by using a hierarchical normal model. Let us first focus on data regarding students from the same school,* $Y_1,\ldots,Y_n$*. Since we lack information distinguishing such students, exchangeability is a reasonable property for* $p(y_1,\ldots,y_n)$*. Thanks to the de Finetti's theorem (we are assuming that the population of individuals in each school is potentially infinite), we have that an equivalent formulation of our information is given by:*

$$\begin{aligned} Y_1,\ldots,Y_n|\gamma &\overset{iid}{\sim} p(y|\gamma) \\ \gamma &\sim p(\gamma) \end{aligned}$$

Now let us consider a model describing our information about the whole hierarchical dataset $(\mathbf{Y}_1,\ldots,\mathbf{Y}_J)$*, where* $\mathbf{Y}_j = \left(Y_{1,j},\ldots,Y_{n_j,j}\right)^T$*. Again it is not reasonable to treat* $Y_{1,j},\ldots,Y_{n_j,j}$ *as independent, since doing so would imply, for example, that the values of* $Y_{1,j},\ldots,Y_{n_j-1,j}$ *would give us no information about* $Y_{n_j,j}$*. However, if all that is known about* $Y_{1,j},\ldots,Y_{n_j,j}$ *is that they are random samples from group* j*, then treating them as exchangeable makes sense:*

$$Y_{1,j},\ldots,Y_{n_j,j}|\gamma_j \overset{iid}{\sim} p(y|\gamma_j).$$

We are left to decide how to represent our information about $\gamma_1, \ldots, \gamma_J$. As before, we do not want to treat these parameters as independent, because doing so would imply that knowing the values of $\gamma_1, \ldots, \gamma_{J-1}$ does not change our information about the value of γ_J. However, if the groups themselves are samples from some larger (infinite) population of groups, then exchangeability of the group-specific parameters might be appropriate. Applying the de Finetti's theorem a second time we get:

$$\gamma_1, \ldots, \gamma_J | \theta \overset{iid}{\sim} p(\gamma_j | \theta)$$

for some unknown random parameter θ. Hence we have:

$$\begin{aligned} Y_{1,j}, \ldots, Y_{n_j,j} | \gamma_j &\overset{iid}{\sim} p(y|\gamma_j) \quad \textbf{\textit{within-group model}} \\ \gamma_1, \ldots, \gamma_J | \theta &\overset{iid}{\sim} p(\gamma_j|\theta) \quad \textbf{\textit{between-group model}} \\ \theta &\sim \pi(\theta) \quad \textbf{\textit{prior distribution}} \end{aligned}$$

In this framework, we assume that the within-group and between-group sampling models are both normal:

$$\begin{aligned} Y_{1,j}, \ldots, Y_{n_j,j} | \theta_j, \sigma^2 &\overset{iid}{\sim} \mathcal{N}(\theta_j, \sigma^2) \text{ for all } j \\ \theta_1, \ldots, \theta_J | \mu, \tau^2 &\overset{iid}{\sim} \mathcal{N}(\mu, \tau^2) \end{aligned}$$

Observe that in this case the mean of the group specific parameter is given by:

$$\mathbb{E}\left[\theta_j \middle| \boldsymbol{y}_j, \mu, \tau^2, \sigma^2\right] = \frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \overline{y}_j + \frac{\frac{1}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \mu$$

In particular, thanks to the exchangeability assumption we have that information is shared across groups: indeed, when n_j is small then $\mathbb{E}\left[\theta_j \middle| \boldsymbol{y}_j, \mu, \tau^2, \sigma^2\right] \approx \mu$ so that the Bayesian estimate is obtained by borrowing strength from the other groups through μ. Finally, we assume that the distributions for the parameters μ, σ^2 and τ^2 are given by the semi-conjugate normal and inverse-gamma prior distributions:

$$\begin{aligned} \frac{1}{\sigma^2} &\sim gamma\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \\ \frac{1}{\tau^2} &\sim gamma\left(\frac{\eta_0}{2}, \frac{\eta_0 \widetilde{\sigma}_0^2}{2}\right) \\ \mu &\sim \mathcal{N}(\mu_0, \gamma_0^2) \end{aligned}$$

```
Y = read.table("school.mathscore.txt")
m = length(unique(Y[,1])) # Number of schools
n = sv = ybar = rep(NA,m)
for(j in 1:m){
    ybar[j] = mean(Y[Y[,1]==j,2]) # Group-specific empircal mean
    sv[j] = var(Y[Y[,1]==j,2]) # Group-specific empircal variance
    n[j] = sum(Y[,1]==j) # Group-specific sample size
}
## Graph of the Data
x11()
par(mfrow=c(1,2),mar=c(3,3,1,1),mgp=c(1.75,.75,0))
hist(ybar,main="",xlab="Sample mean")
plot(n,ybar,xlab="Sample size",ylab="Sample mean")
```

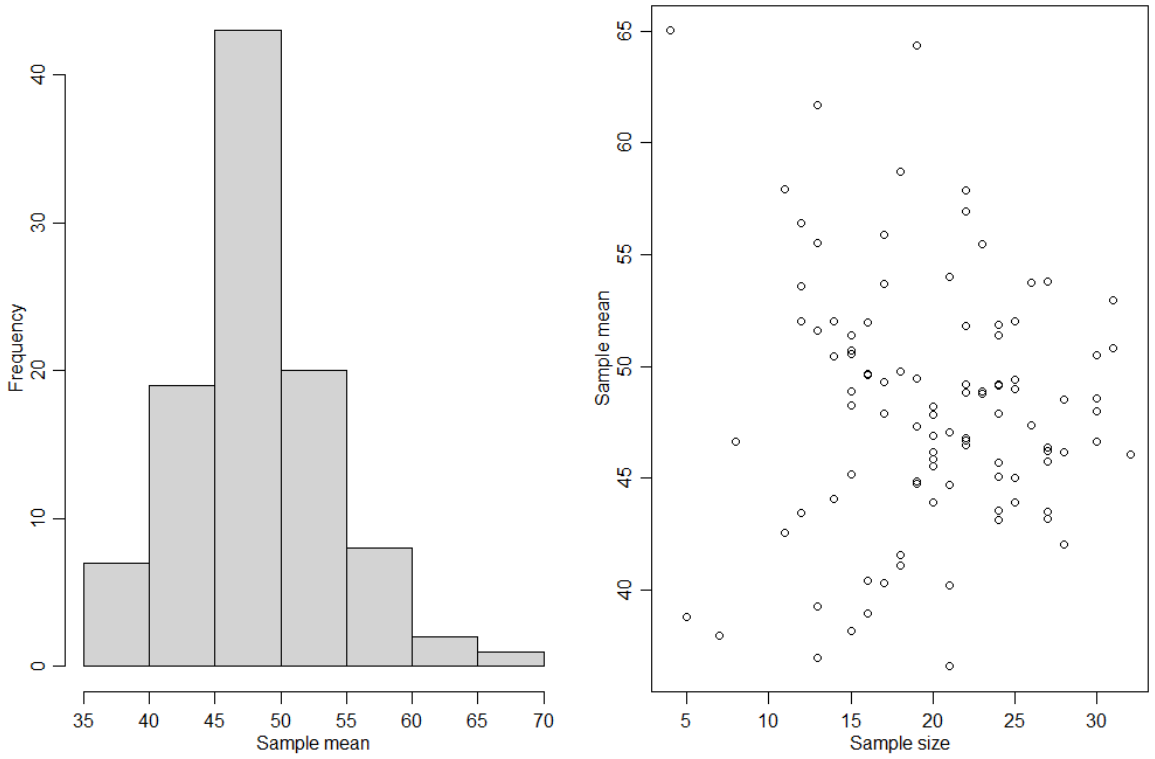
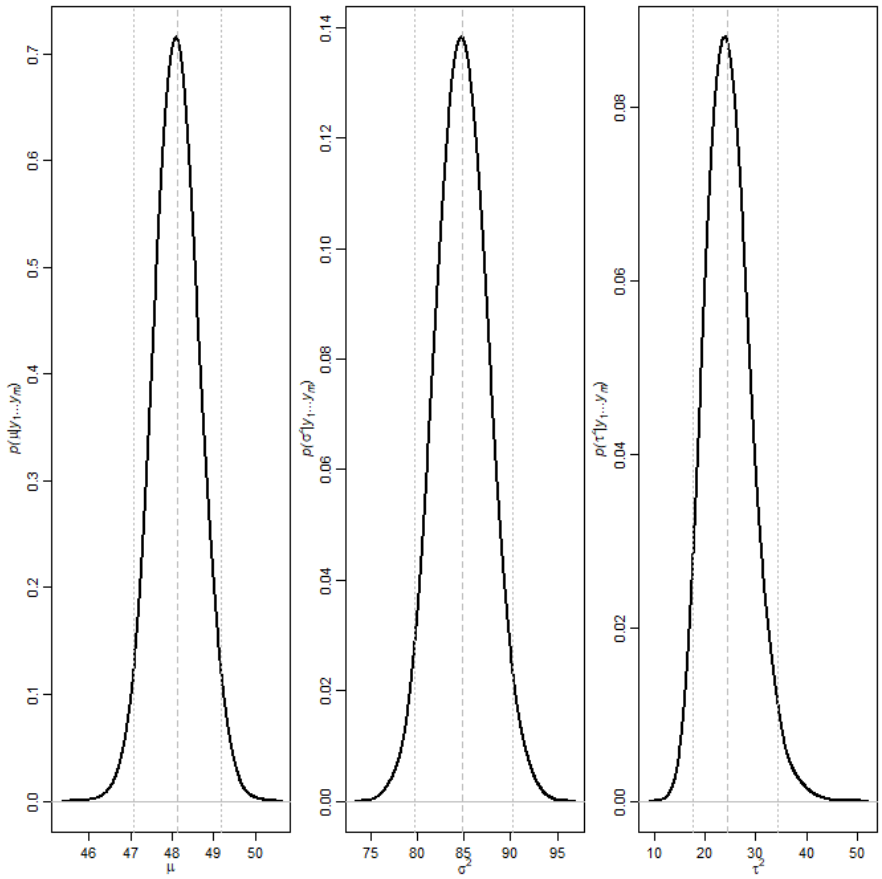


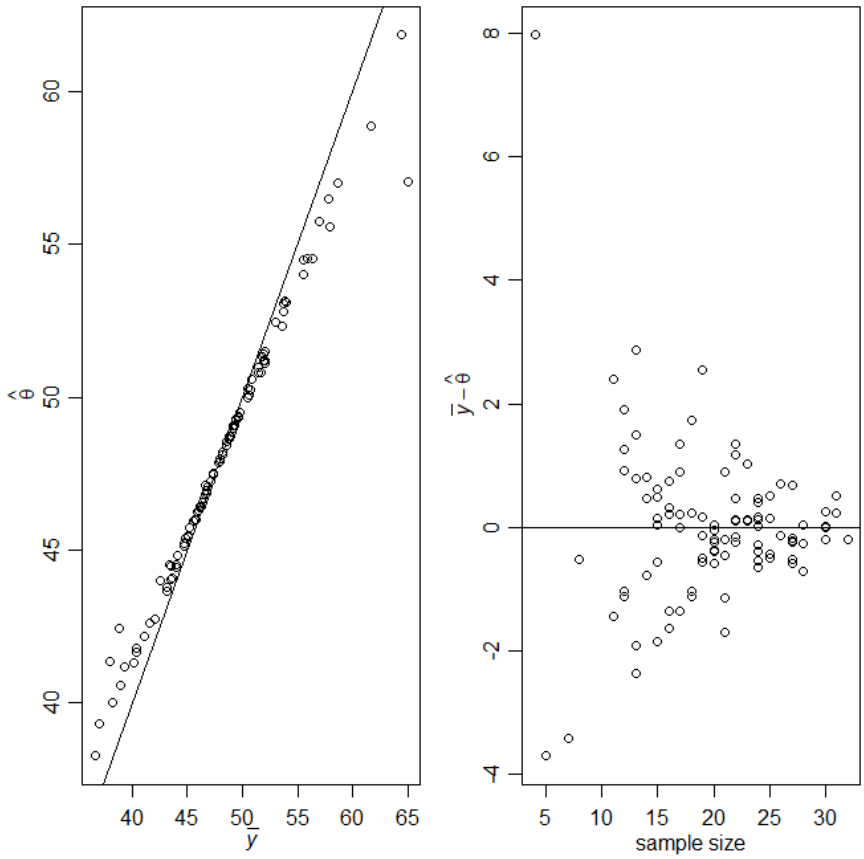
Figure 1: Observe from the plot on the right that very extreme sample averages (very small or very large) tend to be associated with schools with small sample size.

```
## Hyperparameters for the Prior Distributions
nu0 = 1 ; s20 = 100 # Hyperparameters for sigma^2
eta0 = 1 ; t20 = 100 # Hyperparameters for tau^2
mu0 = 50 ; g20 = 25 # Hyperparameters for mu
## MCMC Analysis
### Starting values of the MC
theta = ybar
sigma2 = mean(sv)
mu = mean(theta)
tau2 = var(theta)
### Setup
S = 3000
THETA = matrix(nrow=S,ncol=m)
MST = matrix(nrow=S,ncol=3)
### Gibbs Sampler
for(s in 1:S){
    # Sample new values of the theta's
    for(j in 1:m){
        vtheta = 1/(n[j]/sigma2+1/tau2)
        etheta = vtheta*(ybar[j]*n[j]/sigma2+mu/tau2)
        theta[j] = rnorm(1,etheta,sqrt(vtheta))
    }
    # Sample new value of sigma2
    nun = nu0+sum(n)
    ss = nu0*s20
    for(j in 1:m){
```

```
        ss<-ss+sum((Y[Y[,1]==j,2]-theta[j])^2)
    }
    sigma2 = 1/rgamma(1,nun/2,ss/2)
    # Sample new value of mu
    vmu = 1/(m/tau2+1/g20)
    emu = vmu*(m*mean(theta)/tau2 + mu0/g20)
    mu = rnorm(1,emu,sqrt(vmu))
    # Sample new value of tau2
    etam = eta0+m
    ss = eta0*t20 + sum((theta-mu)^2)
    tau2 = 1/rgamma(1,etam/2,ss/2)
    # Store results
    THETA[s,] = theta
    MST[s,] = c(mu,sigma2,tau2)
}
## Marginal posterior densities of mu, sigma^2 and tau^2
x11()
par(mfrow=c(1,3),mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
plot(density(MST[,1],adj=2),xlab=expression(mu),main="",lwd=2, ylab=
    expression(paste(italic("p("),mu,"|",italic(y[1]),"...",italic(y[m]),
    ")")))
abline(v=quantile(MST[,1],c(.025,.5,.975)),col="gray",lty=c(3,2,3))
plot(density(MST[,2],adj=2),xlab=expression(sigma^2),main="", lwd=2,
ylab=expression(paste(italic("p("),sigma^2,"|",italic(y[1]),"...",italic(
    y[m]),")")))
abline(v=quantile(MST[,2],c(.025,.5,.975)),col="gray",lty=c(3,2,3))
plot(density(MST[,3],adj=2),xlab=expression(tau^2),main="",lwd=2,
ylab=expression(paste(italic("p("),tau^2,"|",italic(y[1]),"...",italic(y[
    m]),")")))
abline(v=quantile(MST[,3],c(.025,.5,.975)),col="gray",lty=c(3,2,3))
```




```
## Shrinkage Effect
### The Bayesian estimate of the group-specific parameters theta_j (given
    the other parameters) is a convex linear combination of the
    frequentist group estimate ybar_j and of mu (the prior mean of all
    theta_j). Therefore, the Bayesian estimate is pulled a bit away from
    ybar_j towards mu by an amount depending on n_j: this is called
    shrinkage. In particular, if n_j is large then the empirical mean is a
    good estimate also from the Bayesian viewpoint and there is no need
    to borrow info from the rest of the groups whereas if it is small then
    we adjust it so that the Bayesian estimate is closer to mu.
x11()
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
theta.hat = apply(THETA,2,mean)
plot(ybar,theta.hat,xlab=expression(bar(italic(y))),ylab=expression(hat(
    theta)))
abline(0,1)
plot(n,ybar-theta.hat,ylab=expression( bar(italic(y))-hat(theta) ),xlab="
    sample size")
abline(h=0)
```



4.1 Linear Mixed Effect Models

We begin with the basic structure of models for normally distributed data that involve group-specific parameters. Let Y_{ij} be a random variable corresponding to an observation. The double subscripts may indicate that this observation is the j-th measurement made on the i-th experimental unit. For instance, there may be repeated observations on different units over time, and Y_{ij} is the measurement made at time j on unit i. Alternatively, there may be multiple units clustered in different groups, and Y_{ij} is the measurement regarding the j-th unit in the i-th group, as it will be the case of the example below. In addition to this structure, there could be vectors of predictor variables associated with each group or with different units within the groups. In particular, let us allow for unit-level covariates: $\boldsymbol{x}_{ij}, i = 1, \ldots, n_j$ (where n_j is the number of units in group j) is a p-dimensional vector of covariates for observation i in group j. We use an ordinary regression model to describe within-group heterogeneity of observations, then describe between-group heterogeneity using a sampling model for the group-specific regression parameters. Expressed symbolically, our within-group sampling model is given by:

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j + \epsilon_{ij}, \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \ i = 1, \ldots, n_j$$

for any fixed $j = 1, \ldots, J$. Parameter β_j is p-dimensional and represents the group-specific regression coefficients. Expressing $Y_{1j}, \ldots, Y_{n_j}$ as a vector $\boldsymbol{Y}_j$ and combining $\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{n_j j}$ into an $n_j \times p$ matrix X_j the within-groups model is given by:

$$\boldsymbol{Y}_j | \boldsymbol{\beta}_j, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(X_j \boldsymbol{\beta}_j, \sigma^2 \mathbb{I}_{n_j}\right), \ j = 1, \ldots, J \tag{13}$$

For the between-group sampling model of the group-specific regression parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$, where J is the total number of groups, if we have no prior information distinguishing the different groups, we can model them as being exchangeable. In particular, in this framework we describe the across-group

heterogeneity with a multivariate normal model:

$$\boldsymbol{\beta}_j|\boldsymbol{\theta},\Sigma \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta},\Sigma), j=1,\ldots,J \tag{14}$$

$$\theta,\Sigma \sim \pi(\theta)\times\pi(\Sigma) \tag{15}$$

where Σ is a $p\times p$ covariance matrix and θ is a p-dimensional vector representing the *average effect* of covariates. According to the specific application and model, the marginal prior of the group-specific parameters might be different, though exchangeable.
The hierarchical regression model (13)-(14) is also called a **linear mixed effects model**. This name is motivated by an alternative parameterization of the previous equations. Indeed, we can rewrite the between-group sampling model as:

$$\boldsymbol{\beta}_j = \boldsymbol{\theta}+\boldsymbol{\gamma}_j, \text{ with } \boldsymbol{\gamma}_j \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0},\Sigma).$$

Plugging this expression into the within-group regression model (13), we get:

$$Y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j+\epsilon_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\theta}+\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}_j+\epsilon_{ij}, \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma^2\right) \tag{16}$$

In this parameterization $\boldsymbol{\theta}$ is referred to as a **fixed effect** as it is constant across groups, whereas $\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_J$ are called **random effects** as they are group-specific can be explained since the regression model contains both fixed and random effects.
Although in (16) the regressors corresponding to the fixed and random effects are the same, this is not compulsory. A more general model is

$$Y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\theta}+z_{ij}^T\boldsymbol{\gamma}_j+\epsilon_{ij}, \tag{17}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ could be vectors of different lengths which may or may not contain overlapping variables. In particular, $\boldsymbol{x}_{ij}$ might contain regressors that are group specific, that is, constant across all observations in the same group. Such variables are not generally included in $\boldsymbol{z}_{ij}$, as there would be no information in the data with which to estimate the corresponding group-specific regression coefficients. The population distribution is such that

$$\boldsymbol{\gamma}_j|\Sigma \overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0},\Sigma),\ j=1,\ldots,J \tag{18}$$

To conclude the definition of the model (17)-(18), we need to specify the prior distribution for the common parameters; typically we assume that
$\pi\left(\boldsymbol{\theta},\Sigma,\sigma^2\right) = \pi(\boldsymbol{\theta})\,\pi(\Sigma)\,\pi\left(\sigma^2\right)$ with:

$$\begin{aligned}
\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, L_0)\\
\Sigma &\sim InvWishart\left(S_0^{-1},\eta_0\right)\\
\sigma^2 &\sim InvGamma\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)
\end{aligned}$$

Example 4.2 (see Hoff (2009), Section 11.3)**.** *Previously we have estimated school-specific expected mathscores, as well as how these expected values vary from school to school. We are now interested in examining the relationship between mathscore and another variable, the socio-economic status (SES), which has been calculated from parental income and education levels for each student in the dataset. In particular, we will assume the following model :*

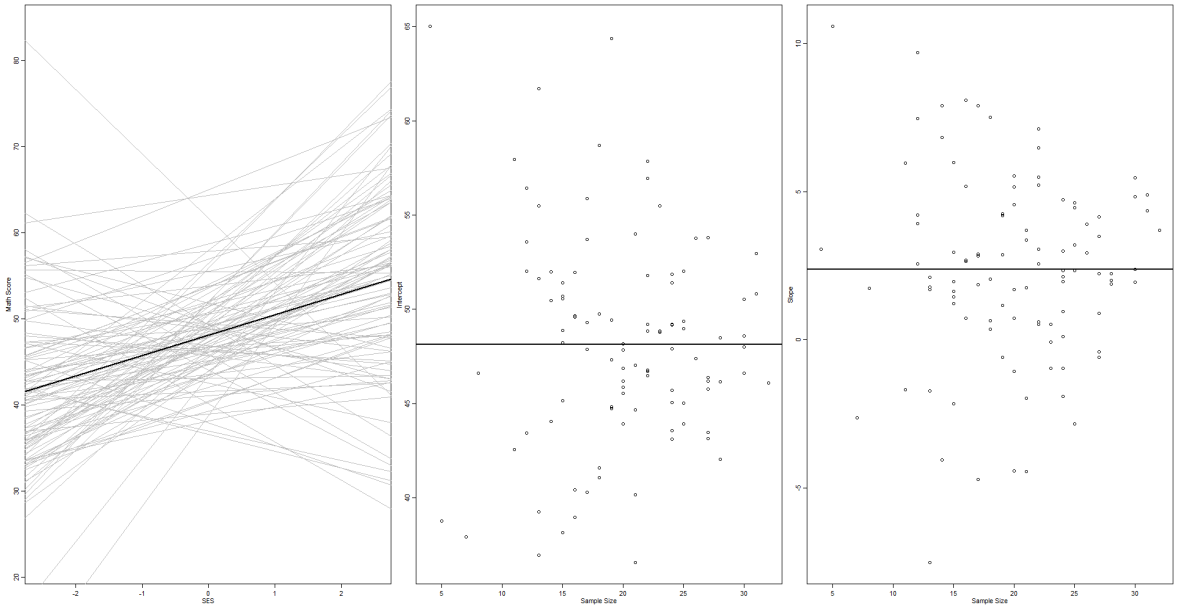
$$Y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j+\epsilon_{ij} = \beta_{1j}+\beta_{2j}SES_{ij}+\epsilon_{ij}, \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}\left(0,\sigma^2\right), i=1,\ldots,n_j, \quad \text{for any } j=1,\ldots,J$$

$$\begin{aligned}
\boldsymbol{\beta}_j|\boldsymbol{\theta},\Sigma &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{\theta},\Sigma)\ \ j=1,\ldots,J\\
\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, L_0)\\
\Sigma &\sim InvWishart\left(S_0^{-1},\eta_0\right)\\
\sigma^2 &\sim inv-gamma\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)
\end{aligned}$$

For what regards the choice of the hyperparameters, we set $\boldsymbol{\mu}_0$ equal to the mean of the least squares estimates within groups, L_0 equal to the sample covariance of such estimates, $\nu_0 = 1$, σ_0^2 equal to the average of the within-groups sample variances, $\eta_0 = p + 1$ and $S_0 = L_0$ (so that the prior for Σ is uninformative). Let us now analyze the R code written to perform the analysis.

```
ldmvnorm = function(X,mu,Sigma,iSigma=solve(Sigma),dSigma=det(Sigma)){
    Y = t(t(X)-mu)
    sum(diag(-1/2*t(Y)%*%Y%*%iSigma)) - 1/2*(prod(dim(X))*log(2*pi) + dim
        (X)[1]*log(dSigma))
}
rmvnorm = function(n,mu,Sigma){
    p = length(mu)
    res = matrix(0,nrow=n,ncol=p)
    if(n>0 & p>0){
        E = matrix(rnorm(n*p),n,p)
        res = t(t(E%*%chol(Sigma)) + c(mu))
    }
    res
}
rwish = function(n,nu0,S0){
    sS0 = chol(S0)
    S = array( dim=c( dim(S0),n ) )
    for(i in 1:n){
        Z = matrix(rnorm(nu0 * dim(S0)[1]), nu0, dim(S0)[1]) %*% sS0
        S[,,i] = t(Z)%*%Z
    }
    S[,,1:n]
}
data = read.table("data_mathscoreSES.txt")
group = data$sch_id
m = length(unique(group))
Y = read.table("school.mathscore.txt")
n = sv = ybar = rep(NA,m)
for(j in 1:m){
    ybar[j] = mean(Y[Y[,1]==j,2])
    sv[j] = var(Y[Y[,1]==j,2])
    n[j] = sum(Y[,1]==j)
}
## Covariates
### X is a list of m = 100 matrices. Each matrix has a number of rows
    equal number of students in the school (included in the sample) and a
    number of columns equal to 2: the first column contains 1's and the
    second one contains the value of the covariate SES (centered)
X = list()
for(j in 1:m){
    xj = data$stu_ses[Y[,1]==j]
    xj = (xj-mean(xj))
    X[[j]] = cbind(rep(1,n[j]),xj)
}
ses_cen = data$stu_ses
## Least Squares Estimate within each Group
S2.LS = BETA.LS = NULL
for(j in 1:m){
    fit = lm(Y[Y[,1]==j,2] ~ -1+X[[j]])
    BETA.LS = rbind(BETA.LS,c(fit$coef))
```

```
    S2.LS = c(S2.LS, summary(fit)$sigma^2)
}
x11()
par(mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
par(mfrow=c(1,3))
## Letf Panel
plot( range(ses_cen),range(Y[,2]),type="n",xlab="SES",ylab="Math Score")
for(j in 1:m){
    abline(BETA.LS[j,1],BETA.LS[j,2],col="gray")
}
BETA.MLS = apply(BETA.LS,2,mean)
abline(BETA.MLS[1],BETA.MLS[2],lwd=2)
## Middle Panel
plot(n,BETA.LS[,1],xlab="Sample Size",ylab="Intercept")
abline(h=BETA.MLS[1],col="black",lwd=2)
## Right Panel
plot(n,BETA.LS[,2],xlab="Sample Size",ylab="Slope")
abline(h= BETA.MLS[2],col="black",lwd=2)
```




```
## Linear Mixed Effect Model
p = dim(X[[1]])[2]
theta = mu0 = apply(BETA.LS,2,mean)
nu0 = 1 ; s2 = s20 = mean(S2.LS)
eta0 = p+2
L0 = matrix(nrow=2,ncol=2)
L0[1,1] = cov(BETA.LS)[1,1]
L0[1,2] = cov(BETA.LS)[1,2]
L0[2,1] = cov(BETA.LS)[2,1]
L0[2,2] = cov(BETA.LS)[2,2]
### Gibbs Sampler Initialization
Sigma = S0 = L0
BETA = BETA.LS
THETA.b = S2.b = NULL
iL0 = solve(L0) ; iSigma = solve(Sigma)
Sigma.ps = matrix(0,p,p)
```

```
SIGMA.PS = NULL
BETA.ps = BETA*0
BETA.pp = NULL
### Gibbs Sampler Cycle
for(s in 1:10000){
    # Update beta_j
    for(j in 1:m){
        Vj = solve(iSigma + t(X[[j]])%*%X[[j]]/s2)
        Ej = Vj%*%(iSigma%*%theta + t(X[[j]])%*%Y[Y[,1]==j,2]/s2)
        BETA[j,] = rmvnorm(1,Ej,Vj)
    }
    # Update theta
    Lm = solve(iL0 + m*iSigma)
    mum = Lm%*%(iL0%*%mu0 + iSigma%*%apply(BETA,2,sum))
    theta = t(rmvnorm(1,mum,Lm))
    # Update Sigma
    mtheta = matrix(theta,m,p,byrow=TRUE)
    iSigma = rwish(1, eta0+m, solve(S0+t(BETA-mtheta)%*%(BETA-mtheta)))
    # Update s2
    RSS = 0
    for(j in 1:m){
        RSS = RSS+sum((Y[Y[,1]==j,2]-X[[j]]%*%BETA[j,])^2)
    }
    s2 = 1/rgamma(1,(nu0+sum(n))/2, (nu0*s20+RSS)/2)
    # Store results
    if(s%%10==0){
        cat(s,s2,"\n")
        S2.b = c(S2.b,s2);THETA.b<-rbind(THETA.b,t(theta))
        Sigma.ps = Sigma.ps+solve(iSigma) ; BETA.ps = BETA.ps+BETA
        SIGMA.PS = rbind(SIGMA.PS,c(solve(iSigma)))
        BETA.pp = rbind(BETA.pp,rmvnorm(1,theta,solve(iSigma)) )
    }
}
x11()
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
plot(density(THETA.b[,2],adj=2),xlim=range(BETA.pp[,2]),main="",xlab="
    Slope Parameter",ylab="Posterior Density",lwd=2)
lines(density(BETA.pp[,2],adj=2),col="gray",lwd=2)
legend(-3,1.0,legend=c(expression(theta[2]),expression(tilde(beta)[2])),
lwd=c(2,2),col=c("black","gray"),bty="n")
BETA.PM = BETA.ps/1000
plot(range(ses_cen),range(Y[,2]),type="n",xlab="SES",ylab="Math Score")
for(j in 1:m){
    abline(BETA.PM[j,1],BETA.PM[j,2],col="gray")
}
abline(mean(THETA.b[,1]),mean(THETA.b[,2]),lwd=2)
```

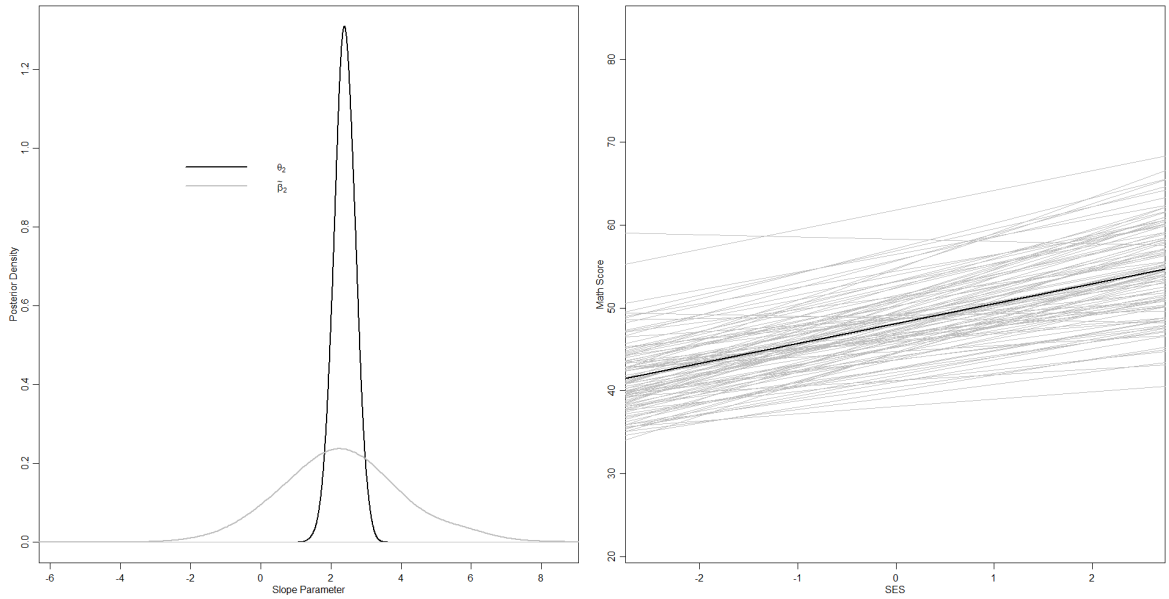


Figure 2: Observe that there is a shrinkage of the 'extreme' frequentist regression lines towards the across-group average line (black line).

5 Model Assessment

The material from Sections 5.1 and 5.2 is taken from Chapter 10 in Rosner et al. (2021).

Until now, our primary inferential goals have been estimation and prediction, and we have generally assumed one model for each data set. For the most part, we have not explored the possibility that there could or would be many alternative models for consideration. Moreover, thus far we have (silently) assumed that fitted models were reasonable for the data without making checks. We now consider the topics of model selection and model checking, which are special cases of model assessment.
Very briefly, if we aim at replying the question Model selection is the way to go when we aim at answering at *which model is the best?*, then we should use tools of *model selection*. If the question is *does our model fit the data well enough?*, we should reply using *model checking*.

5.1 Model Selection

There are two main approaches to model selection:

1. we obtain posterior probabilities that each model is *correct* and (typically) select the model with the largest probability (see Section 10.1 in Rosner et al. (2021)); a special case is when comparing models and this can be done using the Bayes factor, i.e., comparing posterior odds versus prior odds. The Bayes factor criterion is equivalent to comparing the marginals of the observations under the two models.

2. Another approach involves the consideration of prediction-based criteria that involve the sum of a measure of goodness of fit and a penalization for overfitting. Examples of these tools are BIC, DIC, or better WAIC, LPML (see Rosner et al., 2021, Section 10.2).

5.1.1 Model Selection Based on Posterior Probabilities

Choice between Two Models
Suppose that we have two competing models for the data:
model M_0 has a sampling density $\boldsymbol{Y}|\theta_0 \sim f_0(\boldsymbol{y}|\theta_0)$ and a prior $\theta_0 \sim g_0, \theta_0 \in \Theta_0$, whereas
model M_1 has a sampling density $\boldsymbol{Y}|\theta_1 \sim f_1(\boldsymbol{y}|\theta_1)$ and a prior $\theta_1 \sim g_1, \theta_1 \in \Theta_1$. Moreover, let $\pi_0 = \mathbb{P}(M_0)$ and $\pi_1 = \mathbb{P}(M_1) = 1 - \pi_0$ denote the prior probabilities on the two models. We evaluate the posterior probabilities on the two models as follows:

- We first compute the marginal density for the data under the model $M_i, i = 0, 1$:

$$m_i(\boldsymbol{y}) = \int_{\Theta_i} f_i(\boldsymbol{y}|\theta_i)\, g_i(\theta_i)\, d\theta_i, i = 0, 1.$$

  Observe that such quantities express the plausibility of the data $\boldsymbol{y}$ under the model $M_i, i = 0, 1$.

- Then, applying Bayes' theorem, we compute

$$\mathbb{P}(M_0|\boldsymbol{y}) = \frac{\pi_0 m_0(\boldsymbol{y})}{\pi_0 m_0(\boldsymbol{y}) + \pi_1 m_1(\boldsymbol{y})}$$

  and $\mathbb{P}(M_1|\boldsymbol{y}) = 1 - \mathbb{P}(M_0|\boldsymbol{y})$.

As in the case of hypothesis testing, the choice of the model is made by the Bayes factor, i.e., the ratio of posterior odds versus prior odds:

$$BF_{01} = \frac{\frac{\mathbb{P}(M_0|\boldsymbol{y})}{\mathbb{P}(M_1|\boldsymbol{y})}}{\frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)}} = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})}.$$

In general, the computation of the Bayes factor is expensive as it requires two Monte Carlo samples (one from each prior g_0 and g_1). However, if θ_0 and θ_1 have the same dimension and belongs to a

common parameter space Θ, it is straightforward to compute the BF as an integral of one of the two posterior distribution. Indeed, we have:

$$BF_{01} = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} = \frac{\int_\Theta f_0(\boldsymbol{y}|\theta)\, g_0(\theta)\, \mathrm{d}\theta}{\int_\Theta f_1(\boldsymbol{y}|\theta)\, g_1(\theta)\, \mathrm{d}\theta} = \frac{\int_\Theta \frac{f_0(\boldsymbol{y}|\theta)g_0(\theta)}{f_1(\boldsymbol{y}|\theta)g_1(\theta)} f_1(\boldsymbol{y}|\theta)g_1(\theta)d\theta}{\int_\Theta f_1(\boldsymbol{y}|\theta)g_1(\theta)d\theta} = \int_\Theta r(\theta)g_1(\theta|\boldsymbol{y})d\theta$$

where we have defined $r(\theta) = \frac{f_0(\boldsymbol{y}|\theta)g_0(\theta)}{f_1(\boldsymbol{y}|\theta)g_1(\theta)}$. Therefore, if a Markov Chain Monte Carlo sample from $g_1(\theta|\boldsymbol{y})$ is available we can compute BF_{01} as the ergodic mean of the function $r(\theta)$.

Generally, BF computation is heavy from the computational point of view, and often the BF is computed through importance samplers.

See Rosner et al. (2021), Section 10.1.4 for more details.

Choice between Multiple Models
Let us now consider $K = J + 1$ different models $M_0, \ldots, M_J$. We assign the prior probability that each model is *correct*, and then compute the posterior probability that each model is *correct*. We introduce a random parameter m, representing the *correct* index, $\mathfrak{m} = 0, 1, \ldots, J$ and we denote by $\mathbb{P}(\mathfrak{m} = j)$ the prior probability that model M_j is the *correct* one. We typically assume

$$\mathbb{P}(\mathfrak{m} = j) = 1/K \text{ for all } j.$$

We denote by

$$f(\boldsymbol{y}|\theta_j, M_j) \text{ and } \pi(\theta_j|M_j)$$

the likelihood and the prior under model M_j, that is under the event $\mathfrak{m} = j$. Then we compute the posterior probabilities of the models as follows:

(i) first, for any model M_j, we compute the posterior of the parameter vector θ_j via Bayes' theorem as

$$\pi(\theta_j|\boldsymbol{y}, M_j) = \frac{f(\boldsymbol{y}|\theta_j, M_j)\, \pi(\theta_j|M_j)}{m(\boldsymbol{y}|M_j)}$$

where $m(\boldsymbol{y}|M_j) = \int_{\Theta_j} f(\boldsymbol{y}|\theta_j, M_j)\, \pi(\theta_j|M_j)\, d\theta_j$.

(ii) Then we compute

$$\mathbb{P}(\mathfrak{m} = j|\boldsymbol{y}) = \frac{m(\boldsymbol{y}|M_j)\, \mathbb{P}(\mathfrak{m} = j)}{m(\boldsymbol{y})}$$

where $m(\boldsymbol{y}) = \sum_{j=0}^{J} m(\boldsymbol{y}|M_j)\, \mathbb{P}(\mathfrak{m} = j)$.

5.1.2 Model Selection Based on Predictive Information Criteria

We now discuss methods of model selection that are based on how well a model might predict future observations. See Rosner et al. (2021), Section 10.2.

There is a range of predictive model selection criteria that can be regarded as approximations to a particular theoretical criterion. These include the Bayesian information criterion (BIC), the Akaike information criterion (AIC), the deviance information criterion (DIC), the log pseudo marginal likelihood (LPML) criterion, the widely applicable information criterion (WAIC). We particularly recommend the last two, discussing their common motivation and then presenting the different criteria themselves. Given a dataset $\boldsymbol{y} = (y_1, \ldots, y_n)$ and the *new* datapoint $\tilde{y}$, denote by $p(\tilde{y})$ the probability density function for the "true" probability mechanism that generates the observed data and consider a finite set of parametric models $M_0, \ldots, M_J$ for the data at hand. Recall that, if the datapoints are conditionally iid, the (posterior) predictive density based on a candidate model M_j is given by

$$m(\widetilde{y}|\boldsymbol{y}, M_j) = \int_{\Theta_j} f(\widetilde{y}|\theta_j, M_j)\, \pi(\theta_j|\boldsymbol{y}, M_j)\, d\theta_j.$$

If we knew the "true" density $p(\cdot)$, we could use the Kullback-Leibler divergence (KLD) to assess the discrepancy between the "true" density and the predictive density under each candidate model. We would then select the model M_{j^*} that has the smallest KLD from the *true* density. That is, we define an idealized criterion based on KLD called the Kullback-Leibler criterion (KLC):

$$\text{KL}C_j = KLD\left(p(\cdot), m(\cdot|\boldsymbol{y}, M_j)\right) = \int p(\widetilde{y}) \log \frac{p(\widetilde{y})}{m(\widetilde{y}|\boldsymbol{y}, M_j)} d\widetilde{y}$$

Of course, we do not know the true density $p(\cdot)$. Hence, it is not possible to calculate $\text{KL}C_j$, much less minimize it to find model M_{j^*}. The conceptual goal now is to find a surrogate for the KLC that does not depend on unknown $p(\cdot)$ and that can be used to find the model that is "closest" to the true model in the collection under consideration. One such surrogate is given by (minus) the log pointwise predictive density (LPPD):

$$\text{LPPD}_j = \sum_{i=1}^{n} \log m(y_i|\boldsymbol{y}, M_j) \tag{19}$$

More explicitly, we aim to select models with large $\text{LPP}D_j$. One possible problem with this criterion is that observation y_i is being predicted using all of the data, including y_i itself. It is best to avoid this possibility as it may lead to overly optimistic predictions. In these notes, we will see that it is possible to eliminate this problem using two different approaches: (i) add a penalization to $LPPD_j$ or (ii) consider not conditioning to all data, but conditioning to all data but y_i (leave-one-out cross validation).

Log Pseudo-Marginal Likelihood

We adjust overly optimistic predictions in (19) by substituting $m(y_i|\boldsymbol{y}, M_j)$ with $m(y_i|\boldsymbol{y}_{-i}, M_j)$ in the expression of (19), where $\boldsymbol{y}_{-i}$ denotes the data with y_i removed. The method of predicting an observation based on the reduced dataset that does not include it is called *leave-one-out cross-validation.* The corresponding model selection criterion is termed the **log pseudo-marginal likelihood** (LPML) and it is given by:

$$\text{LPML}_j = \sum_{i=1}^{n} \log m(y_i|\boldsymbol{y}_{-i}, M_j) = \sum_{i=1}^{n} \log\left(\text{CPO}_i|M_j\right) \tag{20}$$

where $\text{CPO}_i|M_j = m(y_i|\boldsymbol{y}_{-i}, M_j)$ is called **conditional predictive ordinate** of subject i. We choose the model j with the largest $LPML_j$.

A limitation of this criterion (but only apparent) is that it requires n different Markov Chain Monte Carlo samples (one from each posterior $\pi(\theta_j|\boldsymbol{y}, M_j)$. However, dropping the index j, we observe that:

$$\text{CPO}_i = m(y_i|\boldsymbol{y}_{-i}) = \int_\Theta f(y_i|\theta)\pi(\theta|\boldsymbol{y}_{-i})\,\mathrm{d}\theta = \int_\Theta f(y_i|\theta) \frac{\prod_{l\neq i} f(y_l|\theta)\pi(\theta)}{\int_\Theta \prod_{l\neq i} f(y_l|\theta)\pi(\theta)\mathrm{d}\theta}\mathrm{d}\theta$$

Hence

$$\frac{1}{\text{CPO}_i} = \frac{\int_\Theta \prod_{l\neq i} f(y_l|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \prod_{l=1}^{n} f(y_l|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{\int_\Theta \frac{1}{f(y_i|\theta)} \prod_{l=1}^{n} f(y_l|\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \prod_{l=1}^{n} f(y_l|\theta)\pi(\theta)\mathrm{d}\theta} = \int_\Theta \frac{1}{f(y_i|\theta)}\pi(\theta|\boldsymbol{y})\,\mathrm{d}\theta$$

Therefore, if we have a Markov Chain Monte Carlo sample $\theta^{(1)}, \ldots, \theta^{(M)}$ from the "full" posterior distribution $\pi(\theta|\boldsymbol{y})$ we can approximate $\text{CPO}_i \approx \left[\frac{1}{M}\sum_{t=1}^{M}\left(f\left(y_i|\theta^{(t)}\right)\right)^{-1}\right]^{-1}$.

Widely Applicable Information Criterion

An alternative model-fitting criterion is given by the widely applicable information criterion (WAIC), which is defined by adding a penalization term to the log pointwise predictive density (19):

$$\text{WAIC}_j = \sum_{i=1}^{n} \log m(y_i|\boldsymbol{y}, M_j) - p_{W_j} \tag{21}$$

where $p_{W_j} = \sum_{i=1}^{n} \text{Var}_{\theta_j|\boldsymbol{y}} \log f(y_i|\theta_j, M_j)$. We choose the model j with the largest $LPML_j$.
The predictive density $m(y_i|\boldsymbol{y}, M_j)$ is approximated as $m(y_i|\boldsymbol{y}, M_j) \approx \frac{1}{M} \sum_{t=1}^{M} f\left(y_i \middle| \theta_j^{(t)}, M_j\right)$ where $\theta_j^{(1)}, \ldots, \theta_j^{(M)}$ is a Markov Chain Monte Carlo sample from the posterior of θ_j under the model M_j. All in all, it is possible to prove that under some conditions we have $\text{WAIC}_j \approx LPML_j$ for n large.

According to the probabilistic programming language or the R package, the definition of $WAIC$ is (21) or is -2 times the right hand-side of (21). In this case, we select the model with the lowest value.

5.2 Model Checking

Suppose we have selected a model (or a few) using one or more criteria from the previous sections. We would like to know whether the data are consistent or inconsistent with that model (or those few models). Indeed, the "best" model according to some criteria may still provide a poor fit to the data. More precisely, in considering a model $f(\boldsymbol{y}|\theta_M, M)$ for the data, we now ask whether the data look like they could reasonably have come from that distribution. If the answer is no we should have some concern about this choice of model. A simple approach to answer this question relies on outlier detection via **predictive tail probabilities** (also known as Bayesian predictive p-values):

$$p_i := \min\left\{\mathbb{P}(\widetilde{y} > y_i|\boldsymbol{y}, M), \mathbb{P}(\widetilde{y} < y_i|\boldsymbol{y}, M)\right\}, \ i = 1, \ldots, n$$

Indeed, if the predictive tail probability associated to one observation y_i is too low (typically below 0.05 or 0.1) then such observation can be considered as an outlier for the model M. Therefore, we can say that M is an appropriate model if the number of outliers is small.

Remark. *We could compute the predictive tail probabilities conditioning on the reduced dataset* $\boldsymbol{y}_{-i}$.

For further details see Section 10.3 in Rosner et al. (2021) and the R notebook `Regression_modelchecking_Albert_ex9_2_6`

5.3 Covariate Selection

Our approach to model selection consists in computing the posterior probability mass associated to each model and comparing the models with the largest masses via some predictive goodness-of-fit criterion. However, in the regression framework it is often unfeasible to perform a scan of the entire model space: indeed, the number of models to estimate is 2^k (where k is the number of potential regressors). Therefore, we need an alternative strategy to deal with the model choice problem in this setting. In particular, a lot of attention has been devoted to the development of different regularization methods for simultaneous variable selection and coefficient estimation.

Remark. *The notion of regularization can be meant as imposing additional requirements on the regression solutions in that the more "useful" solutions are preferred over other ones. What is meant by useful depends on the purpose. In particular, in the variable selection framework we have that sparse solutions (i.e., solutions with the redundant coefficients effectively zeroed out) are more desirable.*

5.3.1 A Hierarchical Mixture Model for Variable Selection

Consider an outcome random variable Y that we want to relate to the set of explanatory variables $X_1, \ldots, X_K$ by means of a regression model. The regression framework encompasses a variety of modelling platforms for different types of responses (Gaussian, binary and so on), where the distribution of the response is related to the linear combination of covariates in a way which is specific for the type of outcome. In a GLM, we assume that

$$g[\mathbb{E}(Y|\boldsymbol{x})] = x_1\beta_1 + \cdots + x_K\beta_K.$$

Most often, only some of the available predictors play an important role in explaining the variability of the response, and the goal of the analysis is to identify these variables. We introduce auxiliary variables

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$ such that we are able to model the uncertainty underlying variable selection by using $\boldsymbol{\gamma}$ to augment the parameter space:

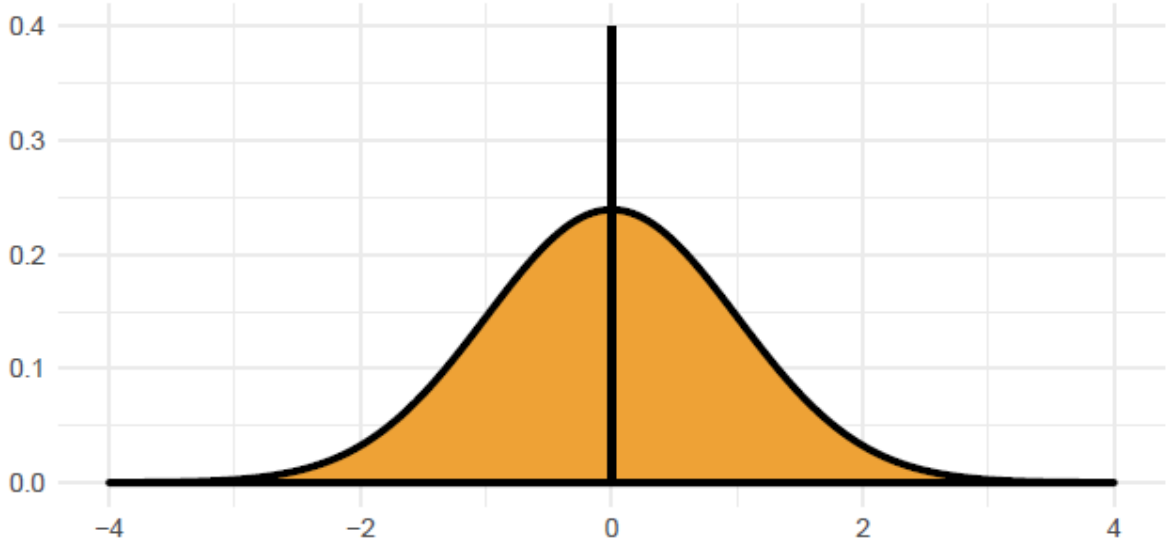
$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}|\boldsymbol{\gamma})\, \pi(\boldsymbol{\gamma})$$

where each $\gamma_j \in \{0, 1\}$. Each *visited* regression model is uniquely characterized by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$ of binary inclusion variables indicating whether or not the variable enters the model. In particular, we assume $\gamma_j \overset{\text{ind}}{\sim} \text{Be}(\theta_j), j = 1, \ldots, K$ and θ_j is the probability that β_j is large enough to justify including X_j in the model.

Spike and Slab Prior
The so-called **spike and slab prior** induces a positive prior probability on the hypothesis $H_0 : \beta_k = 0$. In particular, for any $j = 1, \ldots, K$, such distribution is defined as a mixture of a Dirac measure concentrated at zero and a Gaussian diffuse component:

$$\begin{aligned} \beta_j|\gamma_j &\overset{\text{ind}}{\sim} (1 - \gamma_j)\, \delta_0 + \gamma_j N(0, \sigma^2_{\beta_j}) \\ \gamma_j|\theta_j &\overset{\text{ind}}{\sim} \text{Be}(\theta_j) \\ \theta_j &\overset{\text{ind}}{\sim} \pi(\theta_j) \end{aligned}$$

If $\theta_j = 0.5$ for all j, this corresponds to a uniform prior over all the models.

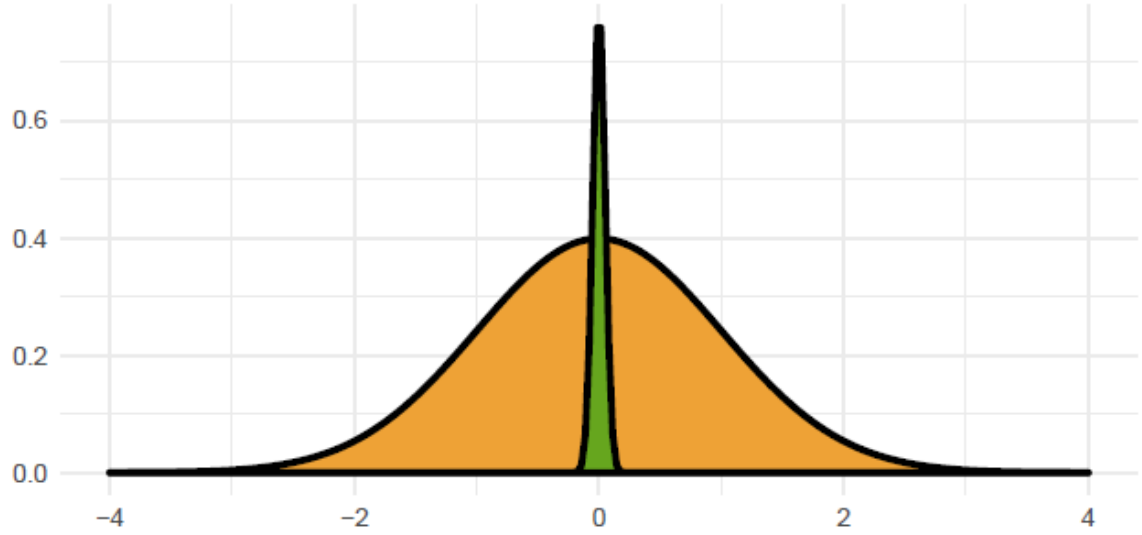


Remark. *A common choice for* $\pi(\theta_j)$ *is* $\mathcal{U}([0, 1])$.

Since most conventional MCMC samplers (e.g., Stan) can only deal with continuous distributions for the parameter, it is often necessary to approximate the spike-and-slab prior with a mixture of two continuous distributions. In particular, one common choice is given by the **stochastic search variable selection prior** (SSVS). In SSVS, the model coefficients β_j are assumed to have a mixture prior of a spike and a slab Gaussian components. The spike element concentrates closely around zero, reflecting the actual absence of the variable in the model. The slab component has a sufficiently large variance to allow the nonzero coefficients to spread over larger values. The degree of separation between the two components is regulated by two tuning parameters τ_j and c_j, where $\tau_j^2 > 0$ is the variance in the spike component and $c_j^2\tau_j^2 > 0$ is the variance in the slab component. To guide the choice of τ_j and c_j, note that the two Gaussian densities intersect at the points $\pm k_j = \tau_j \epsilon_j$, where $\epsilon_j = \sqrt{2\frac{\log(c_j)c_j^2}{c_j^2 - 1}}$. The point k_j can be regarded as a threshold for declaring practical significance in that all coefficients falling into the interval $[-k_j, k_j]$ can be interpreted as "practically zero". Given the

parameter c_j, the variance τ_j^2 can be selected such that the intersection point reflects our perception of practical significance. The mathematical formulation of the SSVS hierarchical prior setup, with binary probit likelihood, is the following:

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta} &\sim \text{Be}\left(\Phi(x_{i1}\beta_1 + \cdots + x_{iK}\beta_K)\right), \quad i = 1, \ldots, n \\ \beta_j | \sigma_j^2 &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma_j^2\right) \; j = 1, \ldots, K \\ \sigma_j^2 | c_j, \tau_j, \gamma_j &\overset{\text{ind}}{\sim} (1 - \gamma_j)\, \delta_{\tau_j^2} + \gamma_j \delta_{c_j^2 \tau_j^2} \\ \gamma_j | \theta_j &\overset{\text{ind}}{\sim} \text{Be}\left(\theta_j\right) j = 1, \ldots, K \\ \theta_j &\overset{\text{ind}}{\sim} \pi\left(\theta_j\right) j = 1, \ldots, K \end{aligned}$$



Criteria for Covariate Selection

As we have introduced earlier, parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$ identifies the m-th *visited* regression model. Indeed $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$ is a vector of binary inclusion variables indicating whether or not the variable enters the model. Let $p\left(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}\right)$ be the likelihood for the n observations under the model indexed by $\boldsymbol{\gamma}$. A natural way to compare models is by inspecting the individual posterior model probabilities:

$$\pi\left(\boldsymbol{\gamma}_0|\boldsymbol{y}\right) = \frac{p\left(\boldsymbol{y}|\boldsymbol{\gamma}_0\right)\pi\left(\boldsymbol{\gamma}_0\right)}{\sum_{\tilde{\boldsymbol{\gamma}}_0} p\left(\boldsymbol{y}|\tilde{\boldsymbol{\gamma}}_0\right)\pi\left(\tilde{\boldsymbol{\gamma}}_0\right)}$$

where $\boldsymbol{\gamma}_0$ and $\tilde{\boldsymbol{\gamma}}_0$ are values that γ may assume and $p\left(\boldsymbol{y}|\boldsymbol{\gamma}_0\right) = \int p\left(\boldsymbol{y}|\boldsymbol{\gamma}_0, \boldsymbol{\beta}\right)\pi\left(d\boldsymbol{\beta}|\boldsymbol{\gamma}_0\right)$ is the marginal density of $\boldsymbol{y}$ in the model indexed by $\boldsymbol{\gamma}_0$. In particular, given an MCMC sample from the posterior distribution $\pi\left(\boldsymbol{\gamma}_0|\boldsymbol{y}\right)$ we can use one of the following strategies to select the covariates:

- **Highest Posterior Probability** (**HPD**). Choose the model with the highest posterior probability:

$$\arg\max_{\boldsymbol{\gamma}_0} \pi\left(\boldsymbol{\gamma}_0|\boldsymbol{y}\right) \simeq \arg\max_{\boldsymbol{\gamma}_0} \frac{1}{m}\sum_{t=1}^{m} \mathbb{I}_{\left(\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}_0\right)},$$

  where $\{\boldsymbol{\gamma}^{(t)}, t = 1, \ldots, m\}$ are the MCMC values from the posterior marginal of $\boldsymbol{\gamma}$.

- **Median Probability Model** (**MPM**). Pick all the covariates with estimated marginal posterior inclusion probability larger than 0.5:

$$j \text{ such that } \pi\left(\gamma_j = 1|\boldsymbol{y}\right) \simeq \frac{1}{m}\sum_{t=1}^{m} \mathbb{I}_{\left(\gamma_j^{(t)} = 1\right)} > 0.5.$$

- **Hard Shrinkage** (**HS**). Pick all the covariates such that 0 does not belong to the posterior marginal credibility interval of the corresponding regressor.

Example 5.1 (see Rockova et al. (2012)). *Rheumatoid arthritis is an autoimmune disease characterized by chronic synovial inflammation and destruction of cartilage and bone in the joints. We are interested in investigating the development of such disease in patients with early manifestations of joint impairment. The data at our disposal consist in basic patient characteristics, serological measurements and patterns of disease involvement for* 681 *different patients. In particular, it is of interest to understand which of the* 12 *observed covariates are potentially associated with the development of rheumatoid arthritis.*

```
reach = read.table("REACH_data.txt", header=T)
reach$gender = reach$gender - 1
X = as.matrix(reach[,1:12])
Y = as.vector(reach[,13])
N = dim(X)[1]
p = dim(X)[2]
## Parameters of the SSVS prior
c_ss = 100
intersect = 0.05
tau_ss = intersect/sqrt(2*log(c_ss)*c_ss^2/(c_ss^2-1))
## Simulate using JAGS
data_JAGS_1 = list(N = N, p = p, Y = Y, X = X, tau_ss = tau_ss, c_ss = c_
    ss) # Data to pass to JAGS
inits = function() {list(beta0 = 0.0, beta = rep(0,p), g = rep(0,p), .RNG
    .seed = 321, .RNG.name = 'base::Wichmann-Hill')} # Initial values of
    the MCMC algorithm
model=jags.model("SSVS_probit.bug",
                 data = data_JAGS_1,
                 n.adapt = 1000,
                 inits = inits,
                 n.chains = 1) # Initialization of the model
update(model,n.iter=1000) # Burn-In = 1000
param = c("beta0", "beta", "g", "mdl") # Posterior parameters to save
nit = 50000 # Number of iterations
thin = 10 # Thinning
output = coda.samples(model = model,
                      variable.names = param,
                      n.iter = nit,
                      thin = thin) # Run the model
output = as.matrix(output)
## MPM Selection Technique
### The estimated posterior inclusion probabilities are the posterior
    means of the gamma variables (called g in the code)
post_g = as.matrix(output[,14:25])
post_mean_g = apply(post_g,2,"mean")
x11()
p2 = data.frame(value = post_mean_g, var = colnames(X)) %>%
     ggplot(aes(y = value, x = var, fill = var)) +
     geom_bar(stat="identity") +
     geom_hline(mapping = aes(yintercept = .5), col = 2, lwd = 1.1) +
     coord_flip() +
     theme_minimal() +
     theme(legend.position="none") +
```

```
    ylab("Posterior Inclusion Probabilities") +
    xlab("")
p2
```

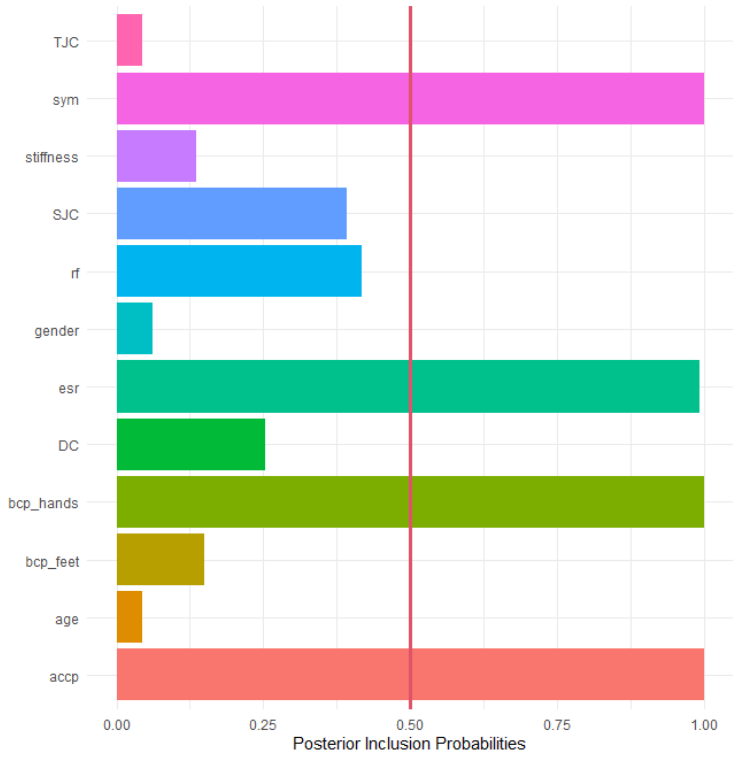


Figure 3: According to the MPM technique we select the covariates sym, esr, bcp_hands and accp.

```
## HPD Selection Technique
### We have represented the model index using a binary encoding: mdl=1+2^
    g[1]+...+2^g[p]
length(unique(output[,"mdl"])) # The chain visited 153 models
visited_models = table(output[,"mdl"])
### We compute the posterior frequency of the visited models
visited_models = table(output[,"mdl"])
### The HPD model is obtained by getting the unique profiles and sorting
    the results
unique_model = unique(post_g, MARGIN = 1)
freq = apply(unique_model, 1, function(b) sum(apply(post_g, MARGIN = 1,
    function(a) all(a == b))))
cbind(unique_model[order(freq,decreasing = T),], sort(freq,decreasing = T
    ))
colnames(X)[as.logical(unique_model[which.max(freq),])] # The selected
    covariates are the same of the previous case.
```

The corresponding JAGS code is given below.

```
# STOCHASTIC SEARCH VARIABLE SELECTION: Probit REGRESSION

model {
    # Likelihood
```

```
    for (i in 1:N) {
        mu[i] <- beta0 + inprod(X[i,], beta)
        z[i] <- phi(mu[i])
        Y[i] ~ dbern(z[i])
    }
    # Tracing the visited model
    for (j in 1:p) {
        TempIndicator[j] <- g[j] * pow(2, j)
    }
    mdl <- 1 + sum(TempIndicator[]) # Index in binary encoding
    c1 <- 1/(pow(tau_ss, 2)) # Reciprocal of the spike variance
    c2 <- c1 / (pow(c_ss, 2)) # Reciprocal of the slab variance
    beta0 ~ dnorm(0, 0.01)
    for(j in 1:p){
        bprior[j] <- 0
        tprior[j] <- equals(g[j],0) * c1 + equals(g[j],1) * c2
        beta[j] ~ dnorm(bprior[j],tprior[j])
        g[j] ~ dbern(0.5) # theta[j] ~ dunif(0,1)
    }
}
```

6 Survival Analysis

We now approach the analysis of data that arise when studying the time until the occurrence of an event. Such data are often called **time-to-event data**, and their analysis is termed **survival** (or **reliability**) **analysis**. For example, we might be interested in the time until a patient diagnosed with leukemia dies or the failure time of an item (i.e. the time until the item *works*). Data in this context are characterized by two features:

- data are realizations of r.v.'s which are positive a.s. and absolutely continuous,
- data can be *censored*.

The interest is typically in studying the time-to-event distribution, for instance, to predict the time-to-event of a new patient joining the study. As one example, a study of the effectiveness of a new anti-cancer drug may focus on the time to death of patients as the primary measurement relevant to the drug's potential benefit. While some patients may have died by the time of the analysis (and their time of death known), others may still be alive. In the latter case, we do not know the time of death but we do know a lower bound for it - the time of analysis. This is also informative and relevant to the goal of the study. Survival analysis is the collection of methods for analyzing such data in which not all individuals have experienced the event of interest and/or we cannot observe the precise time of the event.

6.1 Models for exchangeable Observations

The material in this section is partly based on Rosner et al. (2021), Ch. 11.
We begin by considering situation without covariates, i.e. the case of exchangeable observations.

6.1.1 Survival and Hazard Functions

One key element of survival analysis concerns the so-called **hazard function** or **failure rate**, which we now define. Let T denote a random survival time. Since T is never negative, we may use any probability model for positive random variables to characterize it. Moreover, we will assume that T is absolutely continuous (the most common distributions in this framework are the log-normal distribution, the exponential distribution and the Weibull distribution). Denote by $F(t) = \mathbb{P}(T \leq t)$ the cumulative distribution function for T and call $S(t) = 1 - F(t) = \mathbb{P}(T > t)$ the **survival function**. The hazard function for T is the rate of risk of the event occurring in an instantaneous time interval just beyond t, given that the event has not yet happened:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t)}{\Delta t} \frac{1}{\mathbb{P}(T > t)} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

where $f(t)$ is the density function for T. Alternatively, the hazard function $h(t)$ is the instantaneous propensity to failure/death.

It is well-known that the hazard function identifies the distribution since

$$F(t) = 1 - \exp\{-\int_0^t h(u)du\}, t > 0$$

Moreover, we have

$$P(T > t + s | T > t) = \mathrm{e}^{-\int_t^{t+s} h(u)du}.$$

Hence, we define the following families of distributions:

- increasing failure rate (IFR) distributions, i.e. the hazard $h(t)$ is an increasing function. These distributions represent ageing, and, in fact,

$$P(T > t + s|T > t) < P(T > s) \text{ for all } s, t > 0.$$

- decreasing failure rate (DFR) distributions, i.e. the hazard $h(t)$ is a decreasing function. We also have

$$P(T > t + s|T > t) > P(T > s) \text{ for all } s, t > 0.$$

The exponential distribution is the only absolutely continuous distribution for wich the hazard is constant, i.e. it is DFR and IFR at the same time.

6.1.2 Censoring

We introduce notation that is commonly used in survival analysis/reliability. This notation allows us to write mathematical models for situations in which one is not able to observe the actual time that an event occurred, such as in the examples at the beginning of the chapter.

Right Censoring
We will consider first the case of **right censoring**. This form of censoring occurs when it is not possible to observe the precise time to the event of interest due to another censoring event occurring prior to it. For instance, the censoring event could be the study team losing contact with the patient (*loss to follow-up*) or the time we wish to conduct inference (*end-of-the study time*).
Let C denote the random censoring time. The observed follow-up time Y is the minimum of the time until the event, T, and the censoring time C. With n units, the observed outcome for unit i is $Y_i = \min(T_i, C_i), i = 1, \ldots, n$. In addition to the observed follow-up time, Y_i, we generally also learn whether the unit experienced the event or the censoring time. Data for survival analysis, therefore, consist of a set of follow-up times and a corresponding set of indicators that let us know if each patient's time is the time until the actual event or if it is a censored time. Let δ_i denote the Bernoulli indicator of whether the event occurred at the time for the i-th patient (i.e., the event time is not censored). Thus, δ_i equals 1 if we observe the actual event time for unit i, and it equals 0 if we do not observe the actual event at the end of follow-up. The outcome data in survival analysis with right censored data, therefore, consist of two random quantities, (Y_i, δ_i), for $i = 1, \ldots, n$, where:

$$Y = \min(T, C), \ \delta = \begin{cases} 1, & \text{if } T \leq C \\ 0, & \text{if } T > C. \end{cases}$$

We care about censored data, because we want to include all information in the data set. Even censored observations contain information: a patient still alive 5 years after entering the study tells us something about the treatment, particularly if most patients with this disease tend to die within a year of diagnosis.

Left and Interval Censoring
We now turn to two other kinds of censoring we may encounter: **left censoring** and **interval censoring**. An example of left censoring is the time of infection for a patient diagnosed with pneumonia. We know the infection preceded the time of diagnosis, but we usually do not get to record the actual time the pathogen caused the infection, i.e. $T < t_*$. On the other hand, interval censoring occurs when an event time is both left and right censored, i.e. $T \in (t_*, t^*)$. Interval censoring would occur, for example, if a woman underwent a series of mammography screening examinations every 2 years, starting at age 50, and has a positive mammogram at age 62. We know the cancer reached a detectable level some time between the most recent cancer-positive mammogram and the previous negative one, but we do not get to record the exact time when the tumor became detectable by mammography.

Assumptions for Censoring Mechanism
A common assumption is that censoring is independent of the event of interest, that is, T is independent of C (**independent censoring**). An example where this assumption makes sense is a clinical study that analyzes the study data 5 years after the first patient enters the study. Patients still alive at the time of analysis will be those early entrants with long lives and the later entrants who have not yet been on-study long enough to have experienced the event. As long as there are no trends in the sorts of patients who enter the study over time and long-lived patients are as likely to enter the study early as late in the course of the enrollment period, it seems reasonable to assume that the time of analysis (i.e., the censoring time) is independent of the failure times. Another assumption is that the censoring distribution, say $G(c) = \mathbb{P}(C \leq c)$, does not depend on any of the same parameters as $S(t)$ (**noninformative censoring**). For example, suppose we observe survival in days for a single patient and see the event (uncensored observation) at $y = 25$. Then 25 would be our estimate of the center of the distribution, say the mean or median time of survival. However, if the observation is censored, $(y, \delta) = (25, 0)$, this would make our estimate of the same central measure something larger than 25. Now suppose we have informative censoring so that the time-to-event and censoring distributions have a parameter in common. Specifically, let $T \sim Exp(\theta)$ with mean $\frac{1}{\theta}$ and $C \sim Exp\left(\frac{\theta}{5}\right)$ with mean $\frac{5}{\theta}$. Seeing a censored observation of 25 now makes us think that the median time to event should be about 4 because of the informative censoring. Non-informative censoring is assumed in most survival analysis studies.

6.1.3 Likelihood for Right Censored Data

In the following we will assume that the only censoring mechanism at work in our data set is right censoring; we also assume independent censoring and noninformative censoring. We consider the observations as pairs (Y_i, δ_i) with $Y_i = \min(T_i, C_i)$ he i-th individual's follow-up and $\delta_i = \mathbb{I}(T_i \leq C_i)$ the i-th individual's event indicator. The sampling distribution for the observations (conditional on any model parameters θ) is a product of the contribution from all observations stemming from the usual exchangeable (conditionally independent and identically distributed) model, although some care is needed in deriving this. We begin by noting that the usual model applies to the pairs $(T_i, C_i), i = 1, \ldots, n$ and can be written as:

$$(T_i, C_i) \,|f(\cdot), g(\cdot) \overset{\text{iid}}{\sim} f(\cdot) g(\cdot)$$

where $T_i \sim f(\cdot)$ independent of $C_i \sim g(\cdot)$. However, our observed data are $\{(y_i, \delta_i), i = 1, \ldots, n\}$ and not $\{(t_i, c_i), i = 1, \ldots, n\}$. Therefore, we write (Y_i, δ_i) as the following one-to-one transformation of (T_i, C_i):

$$Y_i = (T_i)^{\delta_i} (C_i)^{1-\delta_i}, \delta_i = \mathbb{I}(T_i \leq C_i)$$

So, the sampling distribution of the data can be written as

$$p((y_1, \delta_1), \ldots, (y_n, \delta_n)|\theta = \{f, g\}) = \prod_{i=1}^{n} p(y_i, \delta_i|\theta).$$

Let us focus first on the case $\delta_i = 1$, i.e. we need to calculate

$$p(y_i, \delta_i = 1|\theta) = P(T_i \in dy_i, C_i > T_i|\theta) = P(T_i \in dy_i, C_i > y_i|\theta) = f(y_i)(1 - G(y_i))$$

where G is the distribution function of C_i (absolutely continuous). The last equality stands because of independence censoring.
Now we compute the contribution of the i-th individual to the likelihood in the case $\delta_i = 0$:

$$p(y_i, \delta_i = 0|\theta) = P(C_i \in dy_i, C_i \leq T_i|\theta) = P(C_i \in dy_i, T_i \geq y_i|\theta) = g(y_i)(1 - F(y_i))$$

where g is the density of C_i. The last equality stands because of independence censoring and because $P(T_i \geq y_i|\theta) = P(T_i > y_i|\theta) = 1 - F(y_i)$. Now,

$$p(y_i, \delta_i|\theta) = \begin{cases} f(y_i)(1 - G(y_i)), & \text{if } \delta_i = 1 \\ g(y_i)(1 - F(y_i)), & \text{if } \delta_i = 0 \end{cases}$$

that is

$$p\left(y_i, \delta_i|\theta\right) \propto \begin{cases} f\left(y_i\right), & \text{if } \delta_i = 1 \\ \left(1 - F\left(y_i\right)\right), & \text{if } \delta_i = 0 \end{cases}$$

because of non informative censoring. Summing up: the sampling distribution (likelihood) is

$$p\left(\left(y_1, \delta_1\right), \ldots, \left(y_n, \delta_n\right)|\theta\right) \propto \prod_{i=1}^{n} \left(f\left(y_i\right)\right)^{\delta_i} \left(1 - F\left(y_i\right)\right)^{1-\delta_i} = \prod_{i=1}^{n} \left(h_i\left(y_i\right)\right)^{\delta_i} S_i\left(y_i\right)$$

Notice here the proportionality (rather than equality) for this sampling distribution, caused by omitting the multiplicative terms in $g\left(\cdot\right)$ and $G\left(\cdot\right)$. This omission would not be possible if there are shared parameters between $f\left(\cdot\right)$ and $g\left(\cdot\right)$, which would be the exceptional case of informative censoring mentioned before.

6.1.4 Parametric Models

In order to perform inference we need to specify the density function $f\left(\cdot\right)$. One possibility is to take $f\left(\cdot\right)$ (and consequently $S\left(\cdot\right)$ and $h\left(\cdot\right)$) to be in a family of distributions indexed by a parameter vector θ (**parametric model**). In this case, given a prior $\pi\left(\theta\right)$ we can resort to MCMC methods to simulate from the posterior $\pi\left(\theta|\boldsymbol{y}, \boldsymbol{\delta}\right)$ and use the simulated sample to perform inference as usual. In particular, some popular targets of inference in this framework are the following characteristics of the conditional distribution of T, given θ:

- The **mean survival time** $\mathbb{E}(T|\theta)$.
- The **median survival time**.
- The **survival probability at a fixed time point**, i.e. $P(T > t|\theta)$.
- The **hazard function** $h(t|\theta)$.

Exponential Model

The simplest survival parametric model assumes the exponential distribution: $T_1, \ldots, T_n|\theta \overset{\text{iid}}{\sim} \text{Exp}\left(\theta\right)$, $\theta > 0$. In this case, the hazard function is $h\left(t\right) = \theta$ for all $t > 0$ so that the likelihood is given by:

$$p\left(\boldsymbol{y}, \boldsymbol{\delta}|\theta\right) \propto \prod_{i=1}^{n} \theta^{\delta_i} e^{-\theta y_i} = \theta^{n_u} e^{-\theta \sum_{i=1}^{n} y_i}$$

where n_u is the number of uncensored observations. It is straightforward to realize that the conjugate prior for θ is given by the gamma distribution:

$$\theta \sim \text{gamma}\left(\alpha, \beta\right) \Longrightarrow \theta|\boldsymbol{y}, \boldsymbol{\delta} \sim \text{gamma}\left(\alpha + n_u, \beta + \sum_{i=1}^{n} y_i\right).$$

Weibull Model

Another popular survival parametric model assumes the Weibull distribution: $T_1, \ldots, T_n|\theta \overset{\text{iid}}{\sim} \text{Weib}\left(\alpha, \lambda\right)$, with $\alpha, \lambda > 0$. The Weibull distribution is a generalization of the exponential distribution: if $T \sim Weib\left(\alpha, \lambda\right)$ then $T^{\alpha} \sim Exp\left(\lambda\right)$. The density function of the Weibull distribution is given by $f\left(t|\alpha, \lambda\right) = \lambda\alpha t^{\alpha-1} \text{e}^{-\lambda t^{\alpha}} \mathbb{1}_{(0,+\infty)}(t)$ and its hazard function is $h\left(t|\alpha, \lambda\right) = \lambda\alpha t^{\alpha-1} \mathbb{1}_{(0,+\infty)}(t)$ so that:

- If $\alpha < 1$, $h(t)$ is monotone decreasing and the distribution is within the DFR family.
- If $\alpha = 1$ $h(t)$ is constant and equal to λ.
- If $\alpha > 1$ $h(t)$ is monotone increasing and the distribution is within the IFR family.

The likelihood for the Weibull model is given by:

$$p\left(\boldsymbol{y}, \boldsymbol{\delta}|\theta\right) \propto (\lambda\alpha)^{n_u} \left(\prod_{i=1}^{n} y_i^{\delta_i}\right)^{\alpha-1} \mathrm{e}^{-\lambda \sum_{i=1}^{n} y_i^{\alpha}}$$

In this case a conjugate prior for the parameters α and λ does not exist, but a popular choice for such distribution is $\pi\left(\alpha, \lambda\right) = \pi\left(\alpha\right)\pi\left(\lambda\right)$ with λ distributed according to a gamma density and α uniformly distributed over some bounded support that includes 1.

6.2 Time-to-Event Regression Models

We now turn to regression analysis of censored data. As in other regression methods, we are interested in discovering or determining associations between explanatory variables and time until the occurrence of an event when that time may be censored. Therefore, in this case the data at our disposal will be given by triplets of the form $(y_i, \delta_i, \boldsymbol{x}_i)$ where $\boldsymbol{x}_i$ is a vector of p covariates associated with the i-th unit. Hence $D = \{(y_i, \delta_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$.

6.2.1 Accelerated Failure-Time Regression Models

Parametric regression models for failure-time data generally consider a linear model for the logarithm of the failure times. A commonly used class of parametric regression models for these data is the accelerated failure-time (AFT) model:

$$\log\left(T_i\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma\varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} F_\varepsilon,\ i = 1, \ldots, n, \tag{22}$$

where F_ε is a **known** distribution function, $\sigma > 0$ is the scale parameter (of the error distribution) and $\boldsymbol{\beta}$ are the regression parameters. Often, we use parameterization

$$\tau := \frac{1}{\sigma^2},\ \text{where } \tau \text{ is the precision.}$$

In the rest of the section, we will use both.
In this case, we write

$$T_i \overset{\text{iid}}{\sim} \mathrm{AFT}\left(F_\varepsilon, \boldsymbol{\beta}, \tau|\boldsymbol{x}_i\right),\ i = 1, \ldots, n. \tag{23}$$

We refer to (23) as an *accelerated failure-time model* because of the following property. Let $W_\sigma := \mathrm{e}^{\sigma\varepsilon}$ so that, from (22), we have $T = \mathrm{e}^{\boldsymbol{x}^T\boldsymbol{\beta}}W_\sigma$ and, for any $t > 0$,

$$F_T\left(t\right) = \mathbb{P}\left(T \leq t\right) = \mathbb{P}\left(\mathrm{e}^{\boldsymbol{x}^T\boldsymbol{\beta}}W_\sigma \leq t\right) = \mathbb{P}\left(W_\sigma \leq t\mathrm{e}^{-\boldsymbol{x}^T\boldsymbol{\beta}}\right) = F_{W_\sigma}\left(t\ \mathrm{e}^{-\boldsymbol{x}^T\boldsymbol{\beta}}\right)$$

We see, then, that the effect of the covariates $\boldsymbol{x}$ on the survival probability is to rescale the time by a factor of $\mathrm{e}^{-\boldsymbol{x}^T\boldsymbol{\beta}}$.
The distribution of the survival time T in AFT models is determined by the distribution of the random variable ε. In particular, common choices are the following:

- If $F_\varepsilon = \mathcal{N}\left(0, 1\right)$, then $\log T_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2\right)$ (log-normal AFT).
- If $F_\varepsilon\left(u\right) = 1 - \mathrm{e}^{-\mathrm{e}^u}$ (Gumbel distribution), then $T_i \overset{\text{ind}}{\sim} \mathrm{Weib}\left(\frac{1}{\sigma}, \mathrm{e}^{-\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}}\right)$ (Weibull AFT).
- If $F_\varepsilon\left(u\right) = \frac{\mathrm{e}^u}{1+\mathrm{e}^u}$ (logistic distribution), then $F_{W_\sigma}\left(t\right) = \frac{t^{\frac{1}{\sigma}}}{1+t^{\frac{1}{\sigma}}}$.

Typically, the prior of parameters $(\boldsymbol{\beta}, \tau)$ is such that either they are independent, i.e., $\pi\left(\boldsymbol{\beta}, \tau\right) = \pi\left(\boldsymbol{\beta}\right)\pi\left(\tau\right)$, or they are conditionally independent, i.e., $\pi\left(\boldsymbol{\beta}, \tau\right) = \pi\left(\boldsymbol{\beta}|\tau\right)\pi\left(\tau\right)$. The marginal priors are those of the regression context.

In the example below (and very often in survival analysis), we focus on the conditional median survival time of T. In the case of a log-normal AFT model, this functional of the unknown parameters is derived as follows:

$$0.5 = \mathbb{P}\left(T \leq m | \boldsymbol{x}, \boldsymbol{\beta}, \sigma\right) = \mathbb{P}\left(\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}\mathrm{e}^{\sigma\varepsilon} \leq m | \boldsymbol{x}, \boldsymbol{\beta}, \sigma\right) = \mathbb{P}\left(\varepsilon \leq \frac{\log m - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma} | \boldsymbol{x}, \boldsymbol{\beta}, \sigma\right) = \Phi\left(\frac{\log m - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma}\right).$$

Since the median of the standard Gaussian distribution is 0, we have that the (conditional) median survival time of a patient with covariate $\boldsymbol{x}$ is

$$m(\boldsymbol{x}) = \mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}.$$

Hence

$$\mathrm{e}^{\beta_j},\ j = 1, \ldots, p$$

is the ratio of two median survival times, associated to two patients with covariate vectors wich differ only of one unit in the j-th covariate.

Example 6.1 (Example 12.1 in Rosner et al. (2021))**.** *We now study an analysis of* 90 *men with cancer of the larynx. The outcome is months from diagnosis until death or censoring for each of the men, and there are three covariates: the stage of verb—Yr—), and the age at diagnosis (*`Age`*). In the analysis, we consider stage* 1 *the baseline and create indicator variables for each stage ($S_{ij} = 1$ if the i-th man's stage of disease is j). The analysis considers the standardized versions of age (*`sAge`*) and year of diagnosis (*`sYr`*). The resulting regression model, leaving out person i's subscripts for ease of reading, is:*

$$\log T = \beta_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 sAge + \beta_6 sYr + \sigma\varepsilon$$

with $\varepsilon \sim \mathcal{N}(0,1)$. We fit a log-normal AFT model to the data using JAGS.

```
library(rstan)
library(rjags)
library(coda)
library(ggmcmc)
df = read.table("https://www.ics.uci.edu/~wjohnson/BIDA/Ch13/Larynx-
    Cancer-Data.txt", header=T)
colnames(df) <- c("stage", "time", "logtime", "age", "yr", "cens_time", "
    log_cens_time")
## In this dataset c = 0 means that lifetime has been OBSERVED
stage = df$stage
t = df$time
age = df$age
yr = df$yr
c = df$cens_time
is.censored = is.na(t) # Indicator of censoring
c[!is.censored] = t[!is.censored] + 1 # Redefine c[i] corresponding to
    observed lifetime as a value larger that t[i] (this is required by
    JAGS)
## JAGS Model
data <- list(oldn = 90, predn = 16,
             stage = stage, t = t[1:90], predt = rep(NA, 16), age = age,
             yr = yr, is.censored = is.censored[1:90], c = c[1:90])
inits = list(beta=c(0,0,0,0,0,0),tau=1)
jags_model = jags.model(file = "LarynxLogNormal.txt", data = data, n.
    chains = 3, inits = inits)
ndraws = 2000
burnin = 1000
update(jags_model, burnin)
```

```
jags_output = coda.samples(jags_model,
                           variable.names = c('beta','predt'),
                           n.iter=ndraws)
jags_df = ggs(jags_output)
jags_df %>%
    group_by(Parameter) %>%
    summarize(mean = mean(value),
              sd = sd(value),
              `2.5%` = quantile(value, .025),
              median = median(value),
              `97.5%` = quantile(value, .975))
data.out = as.matrix(jags_output)
data.out = data.frame(data.out)
## Output
  Parameter  mean    sd       `2.5%` median `97.5%`
  <fct>      <dbl> <dbl>    <dbl>  <dbl>   <dbl>
1 beta[1]    2.31   0.298    1.77    2.30    2.93
2 beta[2]   −0.258  0.496   −1.20   −0.263   0.728
3 beta[3]   −0.962  0.390   −1.73   −0.953  −0.224
4 beta[4]   −2.02   0.503   −3.03   −2.00   −1.06
5 beta[5]   −0.209  0.162   −0.528  −0.209   0.110
6 beta[6]    0.114  0.178   −0.215   0.108   0.481
data.out = as.matrix(jags_output)
data.out = data.frame(data.out)
## Posterior credibility regions for predicted survivals (last 16 rows of
    the dataset)
predsurv = apply(data.out[,7:22],2,"quantile",prob=c(0.025,0.5,0.975))
print(predsurv)
## Output
       predt.1.  predt.2.  predt.3.  predt.4.  predt.5.
2.5%  0.5192374 0.5656554 0.4381702 0.4837101 0.3614699
50%   0.7255680 0.7935597 0.6229246 0.7031551 0.6548949
97.5% 0.8749097 0.9337333 0.7845750 0.8689590 0.8833228
       predt.6.  predt.7.  predt.8.  predt.9. predt.10.
2.5%  0.4640321 0.2859324 0.3887557 0.2374230 0.3032386
50%   0.7337992 0.5463293 0.6322380 0.4540853 0.5429496
97.5% 0.9127552 0.7935290 0.8402186 0.6899481 0.7692444
      predt.11.  predt.12.  predt.13. predt.14.
2.5%  0.02755937 0.05316213 0.1724221 0.2308567
50%   0.11893190 0.17055018 0.3445962 0.4306953
97.5% 0.33814789 0.37792178 0.5572582 0.6542018
       predt.15.  predt.16.
2.5%  0.04418343 0.07889969
50%   0.18715112 0.25219857
97.5% 0.47300516 0.52604770
```

The corresponding JAGS code is given below.

```
### JAGS Code for AFT Model for the Larynx Cancer Dataset
# Note: the last 16 lines of the data file contain points
# at which we wish to make predictions for some combinations
# of Stage, Age and Yr, i.e. they are not actual data!
model{
    ### Create Standardized Covariates & Means
    for (i in 1:(oldn + predn)) {
        sAge[i] <- (age[i]-mean(age[1:oldn]))/sd(age[1:oldn])
```

```
        sYr[i] <- (yr[i]-mean(yr[1:oldn]))/sd(yr[1:oldn])
        mu[i] <- beta[1] + beta[2]*equals(stage[i],2)
                 + beta[3]*equals(stage[i],3)
                 + beta[4]*equals(stage[i],4)
                 + beta[5]*sAge[i] + beta[6]*sYr[i]
        S[i] <- 1- phi((log(5)-mu[i])*sqrt(tau)) # 5 month survival
            probabilities using covariates in the data
        med[i] <- exp(mu[i]) # Medians corresponding to the covariates in
             the data
    }
    ### Model for Observed Subjects
    for (i in 1:oldn) {
        is.censored[i] ~ dinterval(t[i], c[i]) # The indicator is.
            censored takes the value 1 for censored data, 0 otherwise
        t[i] ~ dlnorm(mu[i], tau)
    }
    # Make Predictions for Combinations of Covariates
    for(i in 1:predn) {
        predt[i] <- 1- plnorm(5,mu[i + oldn], tau) # Survival prob after
            5 months for the "new" patients
    }

    for (i in 1:6) {
        beta[i] ~ dnorm(0,0.001)
        rm[i] <- exp(beta[i]) # RMs for each variable
    }
     tau ~ dgamma(2,2)
    sigma <- sqrt(1/tau)
}
```

7 Spatial Models

Climatology, ecology, environmental health, real estate marketing face the task of analyzing data that are:

- highly multivariate, with covariates and response variables;
- geographically referenced;
- temporally correlated.

Spatial data are usually classified into 3 types:

- Point-referenced data (geostatistical data): $Y(\boldsymbol{s})$ random vector at location $\boldsymbol{s} \in \mathbb{R}^r$; $\boldsymbol{s}$ varies continuously over D, a subset of $\mathbb{R}^r$ that contains a r-dim rectangle of positive volume.
- Areal data: $Y(\boldsymbol{s}), \boldsymbol{s} \in D$, and D is partitioned into a finite number of areal units with well-defined boundaries.
- Point pattern data: D is random, and the index set of D gives the locations of random events that are the spatial point pattern. Ex: $Y(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in D$.

Is there any spatial pattern in data $Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), \ldots, Y(\boldsymbol{s}_n)$?
Spatial pattern suggests that measurements for areal units which are near to each other will tend to take more similar values than those for units far from each other.
Independent measurements for the units $\Leftrightarrow$ no pattern.

7.1 References and Software

References:

- Banerjee et al. (2014): Chapters 1, 4, and 6.
- Sahu (2022): Chapter 6 - point referenced data.
- Lee (2022): This vignette is an updated version of a paper in the Journal of Statistical Software in 2013 by the same author.

Software:

- Stan;
- R packages: `INLA`, `spBayes`, `bmstdr` (from Sahu's textbook), `inlabru` (to smooth running INLA), `geostan`, `sptimer`;
- R packages for areal data: `CARBayes`, `CARBayesST`

7.2 Point-referenced Data

Underlying stochastic process $\{Y(\boldsymbol{s}), \boldsymbol{s} \in D\}$, $D \subset \mathbb{R}^r$, here $\mathbb{R}^2$. We observe the stochastic process at n locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$.
Thus, the data is: $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^T$.
The process is strictly stationary if, for any $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n, \boldsymbol{h}$:

$$\mathcal{L}(Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)) = \mathcal{L}(Y(\boldsymbol{s}_1 + h), \ldots, Y(\boldsymbol{s}_n + h))$$

$\gamma(\boldsymbol{h})$ is the semivariogram if, under $\mathbb{E}(Y(\boldsymbol{s} + \boldsymbol{h}) - Y(\boldsymbol{s})) = 0$,

$$\mathbb{E}((Y(\boldsymbol{s} + \boldsymbol{h}) - Y(\boldsymbol{s}))^2) = \mathrm{Var}(Y(\boldsymbol{s} + \boldsymbol{h}) - Y(\boldsymbol{s})) = 2\gamma(\boldsymbol{h}),$$

a function depending solely on $\boldsymbol{h}$; in this case, $\{Y(\boldsymbol{s})\}$ is called intrinsic stationary.
In general: (weakly) stationary $\Rightarrow$ intrinsic stationary.
Relationship between the semivariogram and the covariance function:

$$\gamma(\boldsymbol{h}) = C(0) - C(\boldsymbol{h}), \Leftrightarrow C(\boldsymbol{h}) = C(0) - \gamma(\boldsymbol{h}) = \lim_{\|u\| \to +\infty} \gamma(\boldsymbol{u}) - \gamma(\boldsymbol{h})$$

The process/model is isotropic if $\gamma(\boldsymbol{h}) = \gamma(\|\boldsymbol{h}\|)$.
A process that is intrinsic stationary and isotropic is called homogeneous.

Semivariograms/Covariances for homogeneous processes:

- Exponential:

$$\gamma(d) = \gamma(\|\boldsymbol{h}\|) = \begin{cases} \tau^2 + \sigma^2(1 - \mathrm{e}^{-\phi d}) & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases}$$

$$C(d) = C(\|\boldsymbol{h}\|) = \begin{cases} \sigma^2 \mathrm{e}^{-\phi d} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

  - Nugget: $\tau^2 = \lim_{d \to 0^+} \gamma(d)$. This represents the non-spatial variability of data.
  - Range: $R = \dfrac{1}{\Phi}$, where Φ is the decay parameter.
  - Sill: $\tau^2 + \sigma^2 = \lim_{d \to +\infty} \gamma(d)$.
  - Partial sill: σ^2.
  - Effective range: R_0, distance for which the correlation $\rho(d) = \mathrm{e}^{-\phi d}$ is negligible.

- Powered exponential:

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \mathrm{e}^{-|\phi d|^p}) & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases}, \quad 0 < p \leq 2$$

$$C(d) = \begin{cases} \sigma^2 \mathrm{e}^{-|\phi d|^p} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

- Gaussian:

$$C(d) = C(\|\boldsymbol{h}\|) = \begin{cases} \sigma^2 \mathrm{e}^{-\phi^2 d^2} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

  Note that, for $d > 0$, $C(d) = \sigma^2 \rho(d)$, where, in this case

$$\rho(d; \Phi) = \mathrm{e}^{-\phi^2 d^2}$$

- Matern:

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left(1 - \frac{(\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} \kappa_\nu(\phi d)\right) & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases}$$

$$C(d) = \begin{cases} \sigma^2 \frac{(\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} \kappa_\nu(\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

  κ_ν is the modified Bessel function of order ν.

$$\rho(d; \Phi, \nu) := \frac{(\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} \kappa_\nu(\phi d) \text{ for } d > 0$$

  - $\nu \to +\infty$: Gaussian covariance.
  - $\nu = \dfrac{1}{2}$: exponential covariance.

7.2.1 A Gaussian Spatial Regression Model

Basic model: $Y(\boldsymbol{s}) = \boldsymbol{x}^T(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$.
The residual is partitioned into 2 pieces: $w(\boldsymbol{s})$: spatial residual, $\varepsilon(\boldsymbol{s})$: non-spatial residual.
$\{w(\boldsymbol{s})\}$ is a spatial Gaussian process, capturing the residual spatial association; its distribution involves the parameters σ^2 anf Φ.
$\{\varepsilon(\boldsymbol{s})\}$ represents uncorrelated pure error terms, or microscale variability, i.e. variability at distances smaller than the smallest interlocation distance in the data; τ^2.
It is an homogeneous model, with $C(d)$ exponential, powered exponential and Matern. These models are also called hierarchical models.

Now we introduce the Gaussian spatial regression model. Given X: $n \times p$ matrix with $\boldsymbol{x}^T(\boldsymbol{s}_i)$ as its rows, and $\boldsymbol{w} = (w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n))^T$, the model is:

$$\begin{aligned}
&\boldsymbol{Y}|\boldsymbol{w}, \boldsymbol{\beta}, \tau^2 \sim \mathcal{N}_n(X\boldsymbol{\beta} + \boldsymbol{w}, \tau^2 I_n) \\
&\boldsymbol{w}|\boldsymbol{\theta} \sim \mathcal{N}_n(\boldsymbol{0}, \Sigma(\theta)),\ \Sigma \text{ with entries } \sigma^2\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \boldsymbol{\theta}) \\
&\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_\beta, \Sigma_\beta) \\
&\boldsymbol{\theta} = (\sigma^2, \Phi, \tau^2) \sim \pi \text{ informative}
\end{aligned}$$

An alternative representation integrates out the random effect $\boldsymbol{w}$:

$$\begin{aligned}
&\boldsymbol{Y}|\boldsymbol{\beta}, \theta \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 H(\Phi) + \tau^2 I_n) \quad H_{ij} = \rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \Phi) \\
&\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_\beta, \Sigma_\beta) \\
&\boldsymbol{\theta} = (\sigma^2, \Phi, \tau^2) \sim \pi \text{ informative}
\end{aligned}$$

Often: σ^2, Φ, τ^2 a priori independent, with:

$$\sigma^2 \sim inv - gamma(\cdot, \cdot) \quad \tau^2 \sim inv - gamma(\cdot, \cdot) \quad \Phi \sim \pi$$

7.2.2 Bayesian Kriging

Bayesian kriging is Bayesian prediction!
We want to predict the response y_0 (one-dim) at a new location $\boldsymbol{s}_0$, associated to a p-dim vector $\boldsymbol{x}_0$ of predictors $\boldsymbol{x}(\boldsymbol{s}_0)$, i.e. we compute the predictive distribution of y_0:

$$\begin{aligned}
p(y_0|\boldsymbol{y}, X, \boldsymbol{x}_0) &= \int p(y_0, \boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}, X, \boldsymbol{x}_0) d\boldsymbol{\beta} d\boldsymbol{\theta} \\
&= \int p(y_0|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_0) p(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}, X) d\boldsymbol{\beta} d\boldsymbol{\theta}
\end{aligned}$$

where we assume that $p(y_0|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{x}_0)$ has a conditional Gaussian distribution arising from the joint multivariate distribution of $(Y_0, \boldsymbol{Y})$, i.e.

$$(y_0, \boldsymbol{y})^T|\boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{N}_{1+n}\left((\boldsymbol{x}_0\ X)^T\boldsymbol{\beta}, \tilde{\Sigma}\right)$$

for some *augmented* covariance matrix $\tilde{\Sigma}$ (augmented from $\sigma^2 H(\Phi) + \tau^2 I_n$).

7.2.3 More Insight on Bayesian Models for Geo-referenced Spatio-temporal Data

- Banerjee et al. (2014): Chapter 11.
- Sahu (2022): Chapter 7.

7.3 Areal Data

Data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ (continuous, binary, count) corresponding to n areal units $S = \{S_1, \ldots, S_n\}$.

$$W = [w_{ij}]_{i,j=1,\ldots,n} \quad \text{proximity matrix}$$

entries in W connect units (in some way), $w_{ii} = 0$.
Tipically: $w_{ij} = 1$ if i and j share some common boundary, $w_{ij} = 0$ otherwise. Alternatively, w_{ij} could reflect the distance between units (a decreasing function of intercentroidal distance between the units); the distance could also be returned to a binary determination.
W is usually symmetric and can be standardized: $\tilde{W} = [\tilde{w}_{ij}]$

$$\tilde{w}_{ij} = w_{ij}/w_{i+}, \quad w_{i+} := \sum_j w_{ij}$$

The row sum is always 1, but $\tilde{W}$ is no more symmetric.
w_{ij} can be viewed as a weight: more weight will be associated with j's closer (in some sense) to i than those farther away from i.

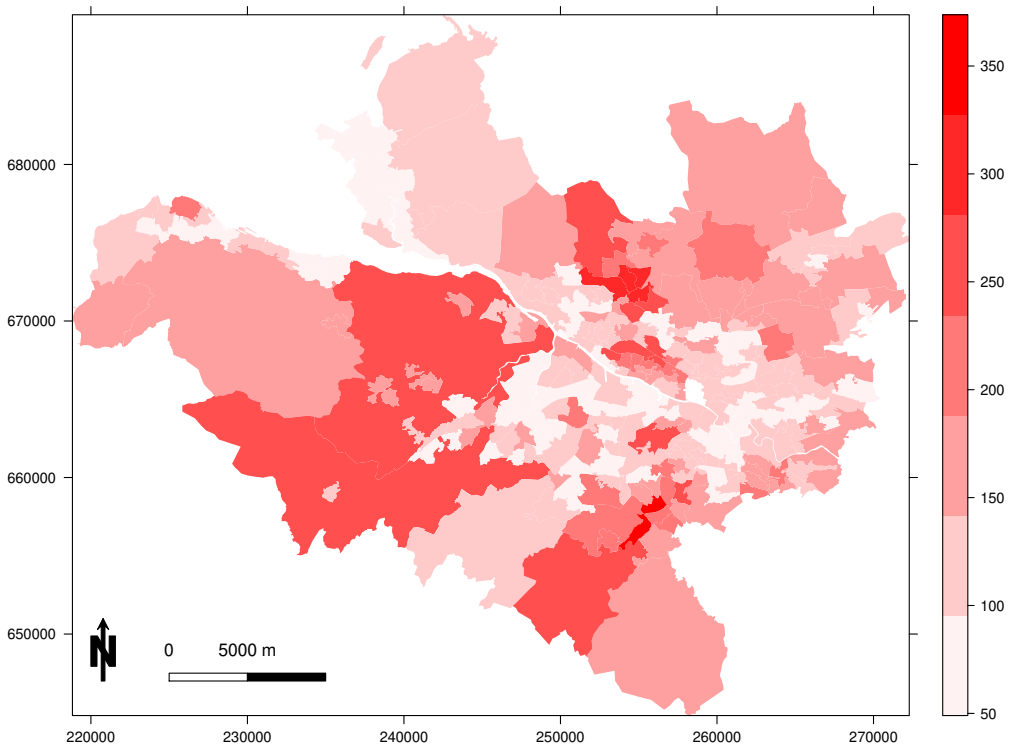


Figure 4: Median property prices in 270 areal units.

Measures of spatial association:

- Moran's I:

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

  It is the analogue of the lagged autocorrelation in time series.
  For large n, under the null hypothesis that Y_i's are i.i.d.: $I \sim \mathcal{N}(-\frac{1}{n-1}, \mathrm{Var}(I))$.
  `moran.test` in `spdep`.

- Geary's C:

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - \bar{Y})^2}{2(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

  For large n, under the null hypothesis that Y_i's are i.i.d.: $C \sim \mathcal{N}(1, \text{Var}(C))$.
  `geary.test` in `spdep`.

Which joint distribution (likelihood) should we choose for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$?
If we give the full conditionals $\mathcal{L}(Y_i|Y_{-i})$, i.e. $p(y_i|y_{-i})$, for all $i = 1, \ldots, n$, do we determine the joint law $p(y_1, \ldots, y_n)$ of $\boldsymbol{Y}$? Yes (under consistency conditions of the full-conditionals), since $p(y_1, \ldots, y_n)$ is determined starting from the product of all $p(y_i|y_{-i})$, but the joint distribution could be improper.

Remark. *if n is large, we do not seek to write down the joint distribution of $\boldsymbol{Y}$; we prefer to work and model exclusively with the n full-conditionals, since they may represent the local behaviour of each Y_i.*

Let ∂_i be a set of neighbors of i. Suppose we specify the full conditionals as

$$p(y_i|y_{-i}) = p(y_i|y_j, j \in \partial_i), i = \ldots, n. \tag{24}$$

If we specify the full-conditionals as in (24), do we uniquely determine the joint law of $(Y_1, \ldots, Y_n)$?
The notion of using local specification as in (24) to determine its joint (global) distribution is referred as Markov Random Field (MRF).
CAR model is an example of MRF such that the corresponding joint distribution is a Gibbs distribution (i.e. it exists, but it can be improper).

7.3.1 Conditionally Autoregressive (CAR) Model

Continuous data Y_i's, Gaussian distributed, it can be extended to data conditionally modeled as the exponential family.
We assume

$$Y_i|y_{-i} \sim \mathcal{N}\left(\sum_{j=1, j \neq i}^{n} b_{ij} y_j, \tau_i^2\right), \quad i = 1, \ldots, n$$

These full conditionals are compatible, and:

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2}\boldsymbol{y}^T D^{-1}(I - B)\boldsymbol{y}\right\},$$

where $B = [b_{ij}]$, $D = diag(\tau_1^2, \ldots, \tau_n^2)$

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{0}, \Sigma_y), \quad \Sigma_y = (I - B)^{-1} D$$

Be careful! $\Sigma_y^{-1} = D^{-1}(I - B)$ must be symmetric.
If $b_{ij} = w_{ij}/w_{i+}$, $\tau_i^2 = \tau^2/w_{i+}$, $W = [w_{ij}]$ proximity matrix,

$$Y_i|y_{-i} \sim \mathcal{N}\left(\frac{\sum_{j=1}^n w_{ij} y_j}{\sum_1^n w_{ij}}, \frac{\tau^2}{\sum_1^n w_{ij}}\right), \ i = 1, \ldots, n,$$

and

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{y}^T (D_w - W)\boldsymbol{y}\right\} \tag{25}$$

where $D_w = diag(w_{1+}, \ldots, w_{n+})$. $(D_w - W)\boldsymbol{1} = \boldsymbol{0} \Rightarrow \Sigma_y^{-1} = D_w - W$ is singular, therefore $\Rightarrow \Sigma_y$ does not exist!
$\Rightarrow p(y_1, \ldots, y_n)$ in (25) is improper and it cannot be used as a model for data.
Some authors suggest to use it (improper as it is!) as a prior in random effects models, as long as the posterior is proper! Do not use it!

7.3.2 Modification of the Intrinsic CAR Model

Redefine $\Sigma_y^{-1} = D_w - \rho W$ and choose ρ to make Σ_y^{-1} non-singular.

- such values for ρ do exist;
- replace W by $\tilde{W} = diag(1/w_{i+})W$; then $\Sigma_y^{-1} = M^{-1}(I - \alpha\tilde{W})$ (M is diagonal) and, if $|\alpha| < 1$, then $I - \alpha\tilde{W}$ is nonsingular.

Under $\Sigma_y^{-1} = D_w - \rho W$ (+ symmetry of W):

$$Y_i|y_{-i} \sim \mathcal{N}\left(\rho\frac{\sum_{j=1}^n w_{ij}y_j}{w_{i+}}, \frac{\tau^2}{w_{i+}}\right)$$

Tipically $\rho \in (0, 1)$:

- $\rho = 0$: $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{\tau^2}{w_{i+}})$, $i = 1, \ldots, n$;
- $\rho = 1$: improperness of $p(y_1, \ldots, y_n)$ in (25) (intrinsic CAR model) .

Prior on ρ in (0,1): a prior on ρ encouraging spatial association puts mass near 1.

7.3.3 GLMM + CAR Prior on the Spatial Random Effects

Exploited in the package `CARBayes`.

- Study region S partitioned into n non-overlapping areal units $S = \{S_1, \ldots, S_n\}$.
- Vector of responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.
- Vector of known offsets $\boldsymbol{O} = (O_1, \ldots, O_n)$.

The spatial pattern in the response is modelled by:

- a matrix of covariates $X = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$, where each $\boldsymbol{x}_k^T = (1, x_{k1}, \ldots, x_{kp})$;
- a set of random effects $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_n)$.

We use the GLMM for the likelihood:

$$\begin{aligned} &Y_k|\mu_k \overset{\text{ind}}{\sim} f(y_k|\mu_k, \nu^2),\ k = 1, \ldots, n \\ &g(\mu_k) = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k \end{aligned} \tag{26}$$

- $f(y_k|\mu_k, \nu^2)$ exponential family [Gaussian, binomial, Poisson];
- $E(Y_k) = \mu_k$;
- ν^2: scale parameter (when needed);
- g: link function [identity, logit, log];
- **Parameters**: $\boldsymbol{\beta}$, ν^2, $\boldsymbol{\Phi}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$, $\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_n)$.

Models for the likelihood:

- **Binomial**: $Y_k \sim Binomial(n_k, \theta_k)$ and
$$\mu_k = ln(\theta_k/(1 - \theta_k)) = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k$$
- **Gaussian**: $Y_k \sim N(\mu_k, \nu^2)$ and $\mu_k = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k$
- **Poisson**: $Y_k \sim Poisson(\mu_k)$ and $ln\,(\mu_k) = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k$

- **ZIP**: $Y_k \sim ZIP(\mu_k, \omega_k)$, zero-inflated Poisson model, used to represent data containing an excess of zeros. This is a mixture of a point mass at zero and a Poisson distribution with mean μ_k.

O_k is the kth offset: they are typically used when data are counts and we want to model rates (instead of counts), and the time (or area) units are different. For instance, suppose our model is Poisson with $E(Y|x) = \mu_x$ but we assume

$$\log \frac{\mu_x}{t_x} = \beta_0 + \beta_1 x \Rightarrow \log \mu_x = \log(t_x) + \beta_0 + \beta_1 x$$

where t_x is the exposure time for covariate x. Then $\log(t_x)$ is the offset.

A priori $\boldsymbol{\beta}$, ν^2, $\boldsymbol{\Phi}$ are independent.

$$\boldsymbol{\beta} \sim \mathcal{N}_{p+1}(\mathbf{0}, 1000 I_{p+1}), \quad \nu^2 \sim inv - gamma\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$\boldsymbol{\Phi} \sim \text{ different priors}$$

Independent prior for the random effects:

$$\Phi_k|\sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \sigma^2 \sim inv - gamma(a, b) \tag{27}$$

The prior for $\boldsymbol{\Phi}$ is exchangeable, since the random effects are conditionally i.i.d. It is appropriate if the covariates included in the model have removed all the spatial structure in the response.
Priors in the `CARBayes` package:

1. intrinsic;
2. BYM models;
3. Leroux et al. (2000);
4. Localised spatial autocorrelation by Lee and Mitchell (2012).

Each prior is a special case of Gaussian Markov Random Field and can be written as:

$$\boldsymbol{\Phi} \sim \mathcal{N}_n(\mathbf{0}, \tau^2 \Sigma^{-1}), \text{ i.e.}$$
$$p(\boldsymbol{\Phi}) \propto \exp\{-\frac{1}{2\tau^2} \boldsymbol{\Phi}^T (D_w - W) \boldsymbol{\Phi}\},$$

with $\Sigma^{-1} = D_w - W$, and W is the proximity matrix.

1. Intrinsic prior:

$$\Phi_k|\Phi_{-k}, W, \tau^2 \sim \mathcal{N}\left(\frac{\sum_{j=1}^n w_{ki}\Phi_i}{w_{k+}}, \frac{\tau^2}{w_{k+}}\right) \tag{28}$$
$$\tau^2 \sim inv - gamma(a = 0.001, b = 0.001)$$

   The conditional expectation of each random effect: $E[\Phi_k|\Phi_{-k}, W, \tau^2]$ = average of the random effects in neighboring areas.
   $\text{Var}[\Phi_k|\Phi_{-k}, W, \tau^2]$= inversely proportional to the number of neighbors.
   Drawbacks:

   (a) the joint prior is improper (but the posterior is proper);

   (b) it represents strong spatial autocorrelation, and produce random effects that are overly smooth.

2. BYM models (most used in practice): replace Φ_k in (26) with $\Phi_k + \theta_k$, where $(\Phi_1, \dots, \Phi_n) \sim$ as in (28) and $(\theta_1, \dots, \theta_n) \sim$ as in (27).
strong + weak spatial correlation.
Identifiability issues: `CARBayes` estimates their sum $re_k = \Phi_k + \theta_k$.

3. Alternative CAR prior, Leroux et al. (2000):

$$\begin{aligned} \Phi_k | \Phi_{-k}, W, \tau^2, \rho &\sim \mathcal{N}\left(\frac{\rho \sum_{j=1}^n w_{ki}\Phi_i}{\rho w_{k+} + 1 - \rho}, \frac{\tau^2}{\rho w_{k+} + 1 - \rho} \right) \\ \tau^2 &\sim inv - gamma(a, b) \\ \rho &\sim \mathcal{U}(0, 1) \end{aligned} \tag{29}$$

Remark. *Values for ρ:*

- $\rho = 0 \Leftrightarrow$ *independence prior.*
- $\rho = 1 \Leftrightarrow$ *intrinsic CAR model prior.*

This model seems the most appealing from both theoretical and practical standpoints.

$$Cor(\Phi_k, \Phi_j | \Phi_{-k,j}, W\rho) = \frac{\rho w_{kj}}{\sqrt{(\rho w_{k+} + 1 - \rho)(\rho w_{j+} + 1 - \rho)}}$$

If $w_{kj} = 0$, i.e. when k and j aren't neighbors, then Φ_k and Φ_j are conditionally independent.
If $w_{kj} = 1$, then their partial autocorrelation is controlled by ρ. This prior is overly simplistic since we are assuming a single global level of spatial smoothing for the set of random effects.

4. Localised spatial autocorrelation by Lee and Mitchell (2012): they should be able to capture localized spatial autocorrelation, including the identification of boundaries in the random effects surface.
Idea: model the elements of W corresponding to geographic adjacent areal units as binary random quantities. Conversely, if (S_k, S_j) do not share a common border, $w_{kj} = 0$.
If w_{kj} is estimated as $1 \Rightarrow \Phi_k$ and Φ_j are spatially correlated, and are smoothed over in the modelling process.
If w_{kj} is estimated as 0, then no smoothing is imparted between Φ_k and Φ_j, as they are modelled as conditionally independent. In this case a boundary is said to exist in the random effects surface between areal units (S_k, S_j).
Goal: the aim is to identify the locations of any boundaries (abrupt step changes) in disease risk surfaces, so the available covariates were used to construct dissimilarity metrics rather than being incorporated into the linear predictor.
They model each w_{kj} as a function of the dissimilarity between areal units (S_k, S_j), because large differences in the response are likely to occur where neighboring populations are very different.
Introduce q non-negative dissimilarity metrics $\mathbf{z}_{kj} = (z_{kj1}, \dots, z_{kjq})$, which could include social or physical factors, such as the absolute difference in smoking rates, or the proportion of the shared border that is blocked by a physical barrier (such as a river or railway line) and cannot be crossed.

Binary model:

$$w_{kj}(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{if } exp\left(-\sum_{i=1}^q z_{kji}\alpha_i\right) \geq 0.5 \text{ and } k \sim j \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_i \overset{\text{ind}}{\sim} \mathcal{U}(0, M_i) \qquad \text{for } i = 1, \dots, q$$

The q regression parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)$ determine the effects of the dissimilarity metrics on $w_{kj} | k \sim j$. For the binary model if $\alpha_i < -\log(0.5)/\max(z_{kij})$, then the i-th dissimilarity metric has not solely identified any boundaries because $\exp(-\alpha_i z_{kij}) > 0.5$ for all $k \sim j$.
Non-binary model:

$$w_{kj}(\boldsymbol{\alpha}) = exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right)$$
$$\alpha_i \overset{\text{iid}}{\sim} Uniform(0, 50) \qquad \text{for } i = 1, \ldots, q$$

The q regression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ determine the effects of the dissimilarity metrics on $w_{kj}|k \sim j$.

`R` functions in the `CARBayes` package:

- `S.glm()`: independent random effects;
- `S.CARbym()`: BYM model (Gaussian likelihood is not allowed);
- `S.CARleroux()`: CAR in Leroux et al. (2000);
- `S.CARdissimilarity()`: local spatial smoothing;
- `S.CARlocalised()`: see the vignette;

7.3.4 Spatio-temporal models

- `CARBayesST` is a dedicated `R` package for spatio-temporal areal unit modelling with conditional autoregressive priors.
- Data on a set of K areal units for N consecutive time periods, yielding a rectangular array of $K \times N$ spatio-temporal observation.
- Observations from geographically close areal units and temporally close time periods tend to have more similar values than units and time periods that are further apart.
- The spatio-temporal structure is modelled via set of autocorrelated random effect.

Fit a generalised linear mixed model to these data, whose general form is

$$\begin{aligned}
&Y_{kt}|\mu_{kt} \sim f\left(y_{kt}|\mu_{kt}, \nu^2\right), \quad \text{for } k = 1, 2, \ldots, K, \ t = 1, \ldots, N \\
&\mu_{kt} = \mathbf{x}_{kt}^T\boldsymbol{\beta} + O_{kt} + \psi_{kt} \\
&\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right)
\end{aligned}$$

where

- O_{kt} is an offset: this value is added to the linear predictor of the target (useful in Poisson regression models, where each case may have different levels of exposure to the event of interest).
- The term ψ_{kt} is a latent component for areal unit k and time period t encompassing one or more sets of spatio-temporally autocorrelated random effects: different spatio-temporal structures are available in the package.

8 Bayesian Nonparametrics

This chapter is partially based on Müller et al. (2015).

In parametric Bayesian inference, we assume that our data are independently drawn from some probability distribution belonging to a parametric family: $Y_1, \ldots, Y_n|\theta \overset{\text{iid}}{\sim} f(\cdot|\theta), \theta \in \Theta \subseteq \mathbb{R}^p$. In this case the unknown parameter θ is finite-dimensional. However, constraining inference to a specific parametric form may limit the scope and type of inferences that can be drawn from such models. In order to allow for more flexibility, we now assume that the population distribution of the data is a random element itself:

$$Y_1, \ldots, Y_n|P \overset{\text{iid}}{\sim} P \tag{30}$$

where P is a random probability distribution on $\mathbb{R}$ (if our datapoints Y_i's are univariate random variables). More in general, P will be a random probability on $\mathbb{R}^p$ if Y_i's are p-dimensional random vectors. The case in (30) is one of the frameworks analyzed by *Bayesian nonparametrics*, when we assume that the parameter is an infinite-dimensional object as in this case (the population probability distribution), when we are interested in making inference on this distribution itself. Other cases concern when the *population density* or the *population hazard* (in survival analysis) are unknown and we assume a prior for them.

We need to introduce:

(i) the space of all probability measures on $\mathbb{R}$ (in case the Y_i's in (30) are unidimensional r.v.'s), denoted by $\mathcal{P} = \mathcal{P}(\mathbb{R})$

(ii) $\mathcal{C}_{\mathcal{P}}$, the σ-field of subsets in $\mathcal{P}$, generated by the open sets in the topology of the weak convergence; $\mathcal{C}_{\mathcal{P}}$ is the smallest σ-field such that the functions $P \mapsto P(B)$ $(B \in \mathcal{B}(\mathbb{R}))$ are measurable.

Definition 8.1 (Random probability measure)**.** *A **random probability measure** is a random variable (element) with values in $\mathcal{P}$ such that $P : (\Omega, \mathcal{F}) \to (\mathcal{P}, \mathcal{C}_{\mathcal{P}})$.*

Hence every determination of P, i.e. $P(\omega)$, is a probability measure on $\mathbb{R}$.

Example 8.1. *Let X be any random variable over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and set $P := \delta_X$ where δ is the Dirac measure. Then P is an example of random probability measure.*

Since we work in a Bayesian framework, we will need to assign a prior distribution for the random probability measure P. In particular, since P is a stochastic process we need to assign the law $\mathcal{L}(P(A_1), \ldots, P(A_k))$ for any finite (and measurable) partition $\{A_1, \ldots, A_k\}$ of $\mathbb{R}$, which is much more difficult than assigning a distribution to a parameter vector. We will see an example of such a prior in the next section.

Summing up, the framework for this section is

$$\begin{aligned} Y_1, \ldots, Y_n|P &\overset{\text{iid}}{\sim} P \\ P &\sim \pi. \end{aligned} \tag{31}$$

P is the random probability measure (r.p.m.) and π represents its prior.

Remark. *The Bayesian estimate for the population distribution of the Y_i's is given by the mean of the posterior distribution for P:*

$$\mathbb{E}[P|Y_1, \ldots, Y_n] = \int_{\mathcal{P}} P \, \pi(dP|Y_1, \ldots, Y_n).$$

As an estimator of a random probability measure, this estimate is a probability measure on $\mathbb{R}$. In particular, observe that such quantity is also equal to the posterior predictive distribution for a new observation Y_{n+1}:

$$\begin{aligned} \mathcal{L}(Y_{n+1}|Y_1, \ldots, Y_n) &= \frac{\int_{\mathcal{P}} \mathcal{L}(Y_1, \ldots, Y_{n+1}|P)\,\pi(dP)}{\int_{\mathcal{P}} \mathcal{L}(Y_1, \ldots, Y_n|P)\,\pi(dP)} = \frac{\int_{\mathcal{P}} \mathcal{L}(Y_{n+1}|P)\mathcal{L}(Y_1, \ldots, Y_n|P)\,\pi(dP)}{\int_{\mathcal{P}} \mathcal{L}(Y_1, \ldots, Y_n|P)\,\pi(dP)} \\ &= \frac{\int_{\mathcal{P}} \mathcal{L}(Y_{n+1}|P)\mathcal{L}(Y_1, \ldots, Y_n|P)\,\pi(dP)}{m(Y_1, \ldots, Y_n)} = \int_{\mathcal{P}} P\pi(dP|Y_1, \ldots, Y_n) \end{aligned}$$

8.1 The Dirichlet Process

One of the most popular Bayesian nonparametric prior is the Dirichlet process prior.

Definition 8.2 (Dirichlet process)**.** *Let α be a finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and let P be a random probability measure. We say that P is a **Dirichlet process** if for any integer k and any finite measurable partition of size k, $\{A_1, \ldots, A_k\}$ of $\mathbb{R}$ we have $(P(A_1), \ldots, P(A_k)) \sim Dir(\alpha(A_1), \ldots, \alpha(A_k))$. In this case we write*

$$P \sim \mathcal{D}_\alpha \text{ or } P \sim DP(a, \alpha_0)$$

with $a := \alpha(\mathbb{R})$ and $\alpha_0(A) := \frac{\alpha(A)}{a}$, $A \in \mathcal{B}(\mathbb{R})$.

Note that notation $\mathcal{D}$ stands for the *Dirichlet measure*, i.e. the distribution of the Dirichlet process. Notation $DP(a, \alpha_0)$, instead, stands for *P is a Dirichlet process*, i.e. a random probability measure.

Remark. *Recall that $(P(A_1), \ldots, P(A_k)) \sim Dir(\alpha(A_1), \ldots, \alpha(A_k))$ means that*

$$P(A_k) = 1 - (P(A_1) + P(A_2) + \ldots + P(A_{k-1}))$$

with $(P(A_1), \ldots, P(A_{k-1}))$ absolutely continuous with density

$$f(x_1, \ldots, x_{k-1}) = \frac{\Gamma\left(\sum_{j=1}^k \alpha(A_j)\right)}{\prod_{j=1}^k \Gamma(\alpha(A_j))} x_1^{\alpha(A_1)-1} \ldots x_{k-1}^{\alpha(A_{k-1})-1} \left(1 - \sum_{j=1}^{k-1} x_j\right)^{\alpha(A_k)-1} \mathbb{I}_{S_{k-1}}(\boldsymbol{x})$$

where S_{k-1} is the $(k-1)$-dimensional simplex, i.e., $S_{k-1} = \{(x_1, \ldots, x_{k-1}) \in \mathbb{R}^{k-1} : 0 \le x_i \le 1, i = 1, \ldots, k-1, \quad 0 \le x_1 + \cdots + x_{k-1} \le 1\}$ is the $k-1$-dimensional simplex.

Proposition 8.1. *Let $P \sim \mathcal{D}_\alpha$. Then for any $A, B \in \mathcal{B}(\mathbb{R})$, we have:*

1. *$P(A) \sim beta\left(\alpha(A), \alpha\left(A^C\right)\right)$.*
2. *$\mathbb{E}[P(A)] = \frac{\alpha(A)}{\alpha(A)+\alpha(A^C)} = \frac{\alpha(A)}{a} = \alpha_0(A)$ and $\mathrm{Var}(P(A)) = \frac{\alpha(A)\alpha\left(A^C\right)}{(a+1)a^2} = \frac{\alpha_0(A)\alpha_0\left(A^C\right)}{a+1}$.*
3. *$\mathbb{E}[P(A)P(B)] = \frac{\alpha(A\cap B)+\alpha(A)\alpha(B)}{a(a+1)}$ and $\mathrm{Cov}(P(A), P(B)) = \frac{\alpha_0(A\cap B)-\alpha_0(A)\alpha_0(B)}{a+1}$.*

Proof. 1. Since $\left\{A, A^C\right\}$ is a finite partition of $\mathbb{R}$ by definition we have that $P(A)$ is absolutely continuous with density
$f(x) = \frac{\Gamma\left(\alpha(A)+\alpha\left(A^C\right)\right)}{\Gamma(\alpha(A))\Gamma(\alpha(A^C))} x^{\alpha(A)-1}(1-x)^{\alpha\left(A^C\right)-1} \mathbb{I}_{S_1}(x)$, that is $P(A) \sim beta\left(\alpha(A), \alpha\left(A^C\right)\right)$.

2. This is a consequence of 1.

3. Suppose first that $A \cap B = \varnothing$. Then $\left\{A, B, (A \cup B)^C\right\}$ is a finite partition of $\mathbb{R}$ so that $(P(A), P(B))$ is absolutely continuous with density

$$f(x_1, x_2) = \frac{\Gamma(a)}{\Gamma(\alpha(A))\Gamma(\alpha(B))\Gamma\left(\alpha\left((A \cup B)^C\right)\right)} x_1^{\alpha(A)-1} x_2^{\alpha(B)-1} (1 - x_1 - x_2)^{\alpha\left((A\cup B)^C\right)-1} \mathbb{I}_{S_2}.$$

Therefore, we compute

$$\begin{aligned}
\mathbb{E}[P(A)P(B)] &= \int_{S_2} x_1 x_2 f(x_1 x_2)\, dx_1 dx_2 = \\
&= \frac{\Gamma(a)}{\Gamma(\alpha(A))\Gamma(\alpha(B))\Gamma\left(\alpha\left((A \cup B)^C\right)\right)} \frac{\Gamma(\alpha(A)+1)\Gamma(\alpha(B)+1)\Gamma\left(\alpha(A \cup B)^C\right)}{\Gamma(a+2)} \\
&= \frac{\alpha(A)\alpha(B)}{a(a+1)}
\end{aligned}$$

and

$$\operatorname{Cov}(P(A), P(B)) = \mathbb{E}[P(A)P(B)] - \mathbb{E}[P(A)]\mathbb{E}[P(B)] = \frac{\alpha(A)\alpha(B)}{a(a+1)} - \alpha_0(A)\alpha_0(B) \quad (32)$$

$$= \frac{\alpha(A)\alpha(B) - \alpha(A)\alpha(B) - a\alpha_0(A)\alpha_0(B)}{a(a+1)} = -\frac{\alpha_0(A)\alpha_0(B)}{a+1}. \quad (33)$$

Let us now consider the case $A \cap B \neq \varnothing$. We have that:

$$\begin{aligned}\operatorname{Cov}(P(A), P(B)) &= \operatorname{Cov}\left(P\left(A \cap B^C\right) + P(A \cap B), P\left(A^C \cap B\right) + P(A \cap B)\right) \\ &= \operatorname{Cov}\left(P\left(A \cap B^C\right), P\left(A^C \cap B\right)\right) + \operatorname{Cov}\left(P\left(A \cap B^C\right), P(A \cap B)\right) \\ &\quad + \operatorname{Cov}\left(P(A \cap B), P\left(A^C \cap B\right)\right) + \operatorname{Var}(P(A \cap B)).\end{aligned}$$

Since the sets $A \cap B^C$, $A^C \cap B$ and $A \cap B$ are pairwise disjoint, it is possible to compute the previous quantity applying (33) to the first three addends, obtaining

$$\operatorname{Cov}(P(A), P(B)) = \frac{\alpha_0(A \cap B) - \alpha_0(A)\alpha_0(B)}{a+1}.$$

□

Proposition 8.2 (Conjugacy of the Dirichlet process)**.** *Let $Y_1, \ldots, Y_n | P \overset{iid}{\sim} P$ and let $P \sim \mathcal{D}_\alpha$. Then the posterior distribution of P is still a Dirichlet process prior, i.e.*

$$P | Y_1, \ldots, Y_n \sim \mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{Y_i}}.$$

Remark. *Observe that a posteriori*

- *the total mass parameters is $a + n$, where $a = \alpha(\mathbb{R})$;*
- *the mean parameter of the Dirichlet process is given by*

$$\frac{\alpha(\cdot) + \sum_{i=1}^n \delta_{Y_i(\cdot)}}{\alpha + n} = \frac{a}{a+n}\alpha_0(\cdot) + \frac{n}{a+n}\left(\frac{1}{n}\sum_{i=1}^n \delta_{Y_i(\cdot)}\right).$$

  Therefore, if n is large we have that the estimate for P is essentially given by the empirical estimate $\frac{1}{n}\sum_{i=1}^n \delta_{Y_i(\cdot)}$.

Definition 8.3 (Support)**.** *Given a probability distribution π on $(\mathcal{P}, \mathcal{C}_\mathcal{P})$ we define **support** of π the smallest closed set $C \subseteq \mathcal{P}$ such that $\pi(C) = 1$.*

Proposition 8.3. *The support of a Dirichlet process is given by*

$$supp(\mathcal{D}_\alpha) = \{p \in \mathcal{P} | supp(p) \subseteq supp(\alpha_0)\},$$

where α_0 is the centering probability measure of P, i.e., $\alpha_0(A) := \dfrac{\alpha(A)}{a}$, $A \in \mathcal{B}(\mathbb{R})$.

Note that, if the support of α_0 is the whole real line, then $\operatorname{supp}(\mathcal{D}_\alpha)$ is the set of all probability measure on $\mathbb{R}$, i.e. the support of $\mathcal{D}_\alpha$ is the full space $\mathcal{P}$. This is a characteristic that all nonparametric prior should have, that is to have full support.

8.1.1 Stick Breaking Construction

The **stick breaking construction** is a constructive definition of the Dirichlet process. Let α be a finite measure on $\mathbb{R}$ and define $a := \alpha(\mathbb{R})$ and $\alpha_0(A) := \frac{\alpha(A)}{a}$, $A \in \mathcal{B}(\mathbb{R})$. Moreover, consider two independent families of random variables $\{Y_i\}$ and $\{\tau_i\}$ such that $Y_1, Y_2, \ldots \overset{\text{iid}}{\sim} \text{beta}(1, a)$ and $\tau_1, \tau_2, \ldots \overset{\text{iid}}{\sim} \alpha_0$. We define

$$P := \sum_{i=1}^{+\infty} V_i \delta_{\tau_i} \tag{34}$$

where

$$\begin{aligned}
V_1 &:= Y_1 \\
V_2 &:= Y_2(1 - Y_1) \\
V_3 &:= Y_3(1 - Y_1)(1 - Y_2) \\
&\vdots \\
V_n &:= Y_n \prod_{j=1}^{n-1}(1 - Y_j) \\
&\vdots
\end{aligned}$$

then we have that $P \sim \mathcal{D}_\alpha$. We do not prove this statement, but we prove that (34) defines a random probability measure.

Proposition 8.4. *We have that*

$$\sum_{j=1}^{+\infty} V_j = 1 \ \textit{a.s.}$$

Proof. We have

$$\begin{aligned}
1 - \sum_{j=1}^{k} V_j &= 1 - Y_1 - Y_2(1 - Y_1) - \ldots - Y_k(1 - Y_1) \cdot \ldots \cdot (1 - Y_{k-1}) \\
&= (1 - Y_1)[1 - Y_2 - Y_3(1 - Y_2) - Y_k(1 - Y_2) \cdot \ldots \cdot (1 - Y_{k-1})] \\
&= (1 - Y_1)(1 - Y_2) \cdot \ldots \cdot (1 - Y_k)
\end{aligned}$$

Therefore, we have that $\mathbb{E}\left[1 - \sum_{j=1}^{k} V_j\right] = \mathbb{E}\left[\prod_{j=1}^{k}(1 - Y_j)\right] \overset{\text{iid}}{=} [\mathbb{E}[1 - Y_1]]^k = \left(\frac{a}{a+1}\right)^k \xrightarrow[k \to +\infty]{} 0$. This implies that

$$\prod_{i=1}^{k}(1 - Y_i) \xrightarrow{a.s.} 0 \text{ as } k \to +\infty.$$

Hence $1 - \sum_{j=1}^{k} V_j \xrightarrow{a.s.} 0$ a.s. as $k \to +\infty$, that means the thesis. □

Remark. *It is clear that the Dirichlet measure selects discrete probability measures.*

8.1.2 Weak Convergence of sequences of Dirichlet Processes

Let $\{\alpha_m\}_{m \geq 1}$ be a sequence of finite measures on $\mathbb{R}$ and set $\alpha_m = a_m \alpha_{0m}$ where $a_m = \alpha_m(\mathbb{R})$. Moreover, suppose that

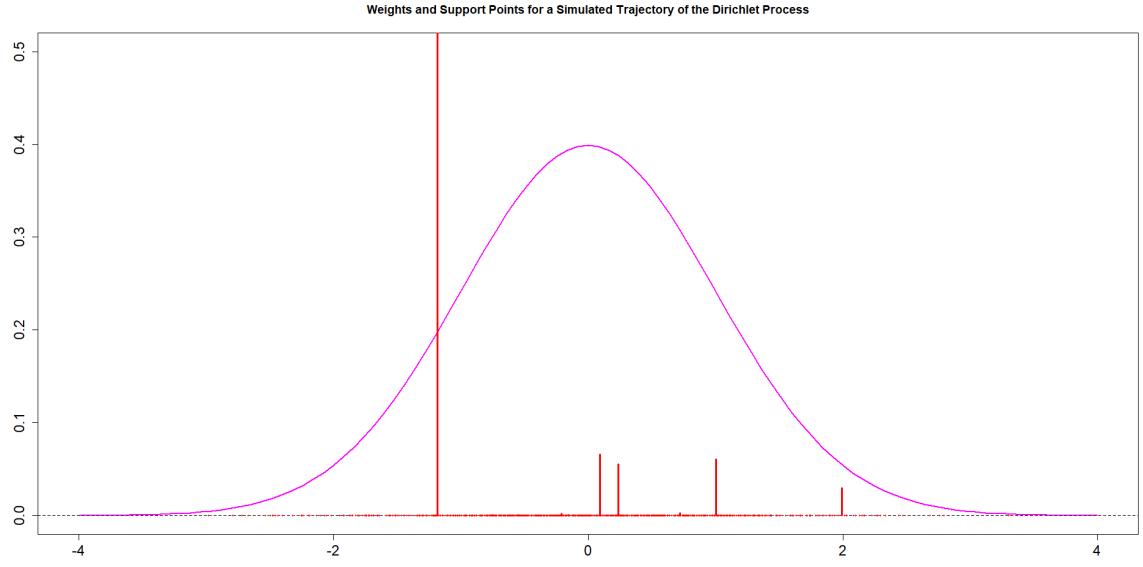
$$\alpha_{0m} \overset{m \to +\infty}{\Longrightarrow} \alpha_0$$

(this is convergence of the associated d.f.'s to any continuity point of the d.f. associated to α_0) where α_0 is a probability measure on $\mathbb{R}$. Then consider a sequence $\{P_m\}_{m \geq 1}$ of random probability measures such that $P_m \sim \mathcal{D}_{\alpha_m} \forall m$. We have that:

- If $a_m \to a$ with $0 < a < +\infty$, then $P_m \xrightarrow{d} P$ as $m \to +\infty$, where $P \sim \mathcal{D}_{a \cdot \alpha_0}$.
- If $a_m \to 0$ then $P_m \xrightarrow{d} \delta_X$ with $X \sim \alpha_0$.
- If $a_m \to +\infty$ then $P_m \xrightarrow{d} \alpha_0$.

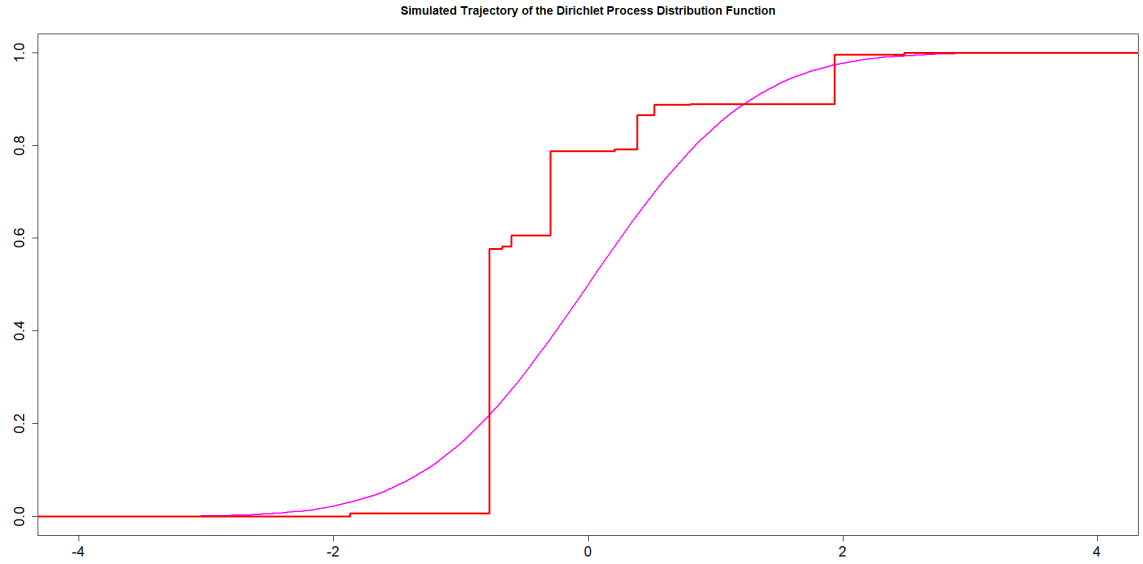
Example 8.2. *We now simulate from a Dirichlet Process defined via the stick-breaking construction and analyze what happens for different values of a.*

```
a = 1 # Total Mass Parameter
M = 1000 # Truncation level of the infinite series
Y = vector(length=M) # Vector of the beta proportions in the Stick-
    Breaking Construction
tau = vector(length=M) # Vector of the support points
V = vector(length=M) # Vector of the weights
## Simulation
Y = rbeta(M,1,a)
tau = rnorm(M,0,1) # Here alpha_0 = N(0,1)
cprod = cumprod(1-Y)
cprod = c(1,cprod[1:M-1])
V = Y*cprod
V = V/sum(V) # Normalization of the weights (needed because of the
    truncation)
## Visualization of the support points and weights compared to alpha_0
x11()
curve(dnorm(x,0,1),from=-4,to=4,col="magenta",lwd=2,ylim=c(0,0.5),xlab="
    ",ylab="",cex.axis=1.5)
abline(h=0,lty=2)
lines(tau,V,"h",lwd=3,col="red")
title("Weights and Support Points for a Simulated Trajectory of the
    Dirichlet Process")
```



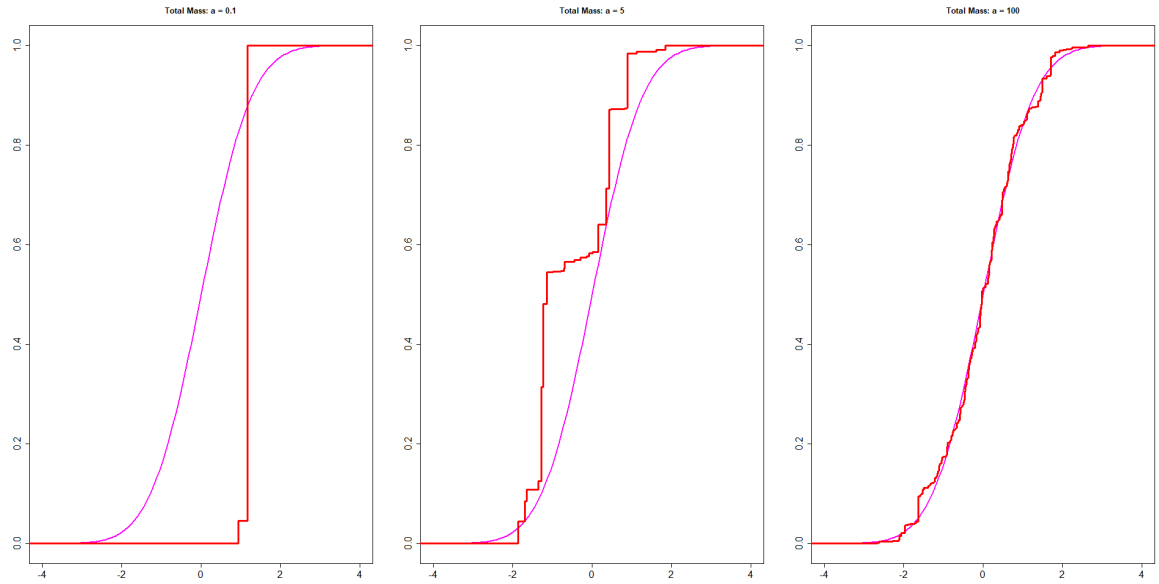

```
## Visualization of the trajectory of the corresponding distribution
    function
x11()
```

```
oth = order(tau)
curve(pnorm(x,0,1),from=-4,to=4,col="magenta",lwd=2,xlab="",ylab="",cex.
    axis=1.5)
lines(c(min(tau)-100,tau[oth],max(tau)+100),c(0,cumsum(V[oth]),1),type="s
    ",col="red",lwd=3)
title("Simulated Trajectory of the Dirichlet Process Distribution
    Function
```




```
## Comparison for 3 Values of a
####
a = 0.1
M = 1000
Y = vector(length=M)
tau = vector(length=M)
V = vector(length=M)
## Simulation
Y = rbeta(M,1,a)
tau = rnorm(M,0,1)
cprod = cumprod(1-Y)
cprod = c(1,cprod[1:M-1])
V = Y*cprod
V = V/sum(V)
x11()
par(mfrow=c(1,3))
oth = order(tau)
curve(pnorm(x,0,1),from=-4,to=4,col="magenta",lwd=2,xlab="",ylab="",cex.
    axis=1.5)
lines(c(min(tau)-100,tau[oth],max(tau)+100),c(0,cumsum(V[oth]),1),type="s
    ",col="red",lwd=3)
title("Total Mass: a = 0.1")
####
a=5
M = 1000
Y = vector(length=M)
tau = vector(length=M)
```

```
V = vector(length=M)
## Simulation
Y = rbeta(M,1,a)
tau = rnorm(M,0,1)
cprod = cumprod(1-Y)
cprod = c(1,cprod[1:M-1])
V = Y*cprod
V = V/sum(V)
oth = order(tau)
curve(pnorm(x,0,1),from=-4,to=4,col="magenta",lwd=2,xlab="",ylab="",cex.
    axis=1.5)
lines(c(min(tau)-100,tau[oth],max(tau)+100),c(0,cumsum(V[oth]),1),type="s
    ",col="red",lwd=3)
title("Total Mass: a = 5")
####
a = 100
M = 1000
Y = vector(length=M)
tau = vector(length=M)
V = vector(length=M)
## Simulation
Y = rbeta(M,1,a)
tau = rnorm(M,0,1)
cprod = cumprod(1-Y)
cprod = c(1,cprod[1:M-1])
V = Y*cprod
V = V/sum(V)
oth = order(tau)
curve(pnorm(x,0,1),from=-4,to=4,col="magenta",lwd=2,xlab="",ylab="",cex.
    axis=1.5)
lines(c(min(tau)-100,tau[oth],max(tau)+100),c(0,cumsum(V[oth]),1),type="s
    ",col="red",lwd=3)
title("Total Mass: a = 100")
```



8.1.3 Marginal Distribution of a Sample from a Dirichlet Process

Consider a random sample

$$\begin{aligned} Y_1, \ldots, Y_n | P &\overset{\text{iid}}{\sim} P \\ P &\sim DP(a, \alpha_0). \end{aligned}$$

A key property of the Dirichlet Process is its discreteness, which implies a positive probability of ties among the Y_i's. This is at the heart of the Polya urn representation for the marginal distribution $\mathcal{L}(Y_1, \ldots, Y_n)$, which specifies such law as a product of sequence of increasing conditionals:

$$\mathcal{L}(Y_1, \ldots, Y_n) = \mathcal{L}(Y_1)\, \mathcal{L}(Y_2|Y_1) \ldots \mathcal{L}(Y_n|Y_1, \ldots, Y_{n-1})$$

Let us now compute such expression explicitly. We have that

$$\mathcal{L}(Y_1) = \int_{\mathcal{P}} \mathcal{L}(Y_1, dP) = \int_{\mathcal{P}} \mathcal{L}(Y_1|P)\, \mathcal{L}(\mathrm{dP}) = \int_{\mathcal{P}} P\mathcal{L}(\mathrm{dP}) = \mathbb{E}(P) = \alpha_0,$$

because $P \sim DP(a, \alpha_0)$ and $\mathbb{E}(P) = \alpha_0$. Note that any marginal distribution of Y_i is α_0. Now, because $Y_1, Y_2, \ldots, Y_n$ (for any n) are conditionally independent, given P, we have

$$\begin{aligned} \mathcal{L}(Y_2|Y_1) &= \int_{\mathcal{P}} \mathcal{L}(Y_2, dP|Y_1) = \int_{\mathcal{P}} \mathcal{L}(Y_2|P, Y_1)\, \mathcal{L}(\mathrm{dP}|Y_1) = \int_{\mathcal{P}} \mathcal{L}(Y_2|P)\, \mathcal{L}(\mathrm{dP}|Y_1) = \int_{\mathcal{P}} P\mathcal{D}_{\alpha+\delta_{Y_1}}(\mathrm{dP}) \\ &= \frac{a}{a+1}\alpha_0 + \frac{1}{a+1}\delta_{Y_1} \\ &\vdots \\ \mathcal{L}(Y_n|Y_1, \ldots, Y_{n-1}) &= \int_{\mathcal{P}} P\mathcal{L}(\mathrm{dP}|Y_1, \ldots, Y_{n-1}) = \frac{a}{a+n-1}\alpha_0 + \frac{n-1}{a+n-1}\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\delta_{Y_i}\right) \end{aligned}$$

Therefore, we have

$$\mathcal{L}(Y_1, \ldots, Y_n) = \alpha_0 \prod_{i=2}^{n} \frac{a\alpha_0 + \sum_{j=1}^{i-1}\delta_{Y_j}}{a+i-1} \tag{35}$$

Remark. *This mechanism of sampling items* $\{1, 2, \ldots, n\}$ *from* (35) *is called generalized Pólya urn.*

Example 8.3. *We now simulate a random sample* $Y_1, \ldots, Y_n | P \overset{iid}{\sim} P, P \sim DP(a, \alpha_0)$ *in two different ways:*

1. *We simulate a Dirichlet process defined via the (truncated) stick-breaking construction, obtaining* P_M *and then we simulate the sample* $Y_1, \ldots, Y_n$ *as iid from* P_M*.*
2. *We simulate the sample* $Y_1, \ldots, Y_n$ *directly via the Pólya urn scheme* (35)*.*

```
## First way
a = 1
M = 100
Y = vector(length=M)
tau = vector(length=M)
V = vector(length=M)
## Simulation
Y = rbeta(M,1,a)
tau = rnorm(M,0,1)
cprod = cumprod(1-Y)
```

```
cprod = c(1,cprod[1:M-1])
V = Y*cprod
V = V/sum(V)
## Simulation of an iid sample from P (for this specific trajectory)
n = 10 # Sample size
theta = vector(length=n) # Simulate values
index = vector(length=n)
for(j in 1:n){
    index[j] = sample(1:M, size=1,prob=V)
    theta[j] = tau[index[j]]
}
k = length(unique(theta)) # The number of unique values in the sample is
    k <= n due to the discreteness of the Dirichlet Process
a = 1
n = 10
theta = vector(length=n)
theta[1] = rnorm(1) # First simulate a value from alpha_0 = N(0,1)
for(j in 2:n){
    w0 = a/(j-1+a) # Probability that we simulate a new observation from
        alpha_0
    w1 = rep(1/(j-1+a),j-1) # Probability that we simulate an old
        observation
    index = sample(0:(j-1),size=1,prob=c(w0,w1))
    if(index==0){
        theta[j] = rnorm(1)
    }
    else{
        theta[j] = theta[index]
    }
}
k = length(unique(theta)) # As before we have that k <= n
```

Random partition model induced by a sample from a Dirichlet process

*Let us now observe that, due to the discreteness of Dirichlet processes, if $(Y_1, Y_2, \ldots, Y_n)$ is a sample from a Dirichlet process with parameter α, we can use the ties among the Y_i's to define a random partition. Let $\{Y_i^*, i = 1, \ldots, k\}$ denote the unique values among the Y_i's and define clusters $S_i = \{$**all** $j = 1, \ldots, n : Y_j = Y_i^*\}$, so that the random partition is given by $\rho_n = \{S_1, \ldots, S_k\}$. It is possible to prove that:*

$$\mathbb{P}(\rho_n = \{S_1, \ldots, S_k\}) = p(n_1, \ldots, n_k) = \frac{\Gamma(a)}{\Gamma(a+n)} a^k \Gamma(n_1) \cdot \ldots \cdot \Gamma(n_k), \quad \textbf{for } n_i \geq 0, n_1 + \cdots + n_k = n \tag{36}$$

where $n_i = |S_i|$. Note that in the formula above the number of clusters k ***is random****. This is also known as Ewens sampling formula.*

The marginal prior for the number of clusters k (induced by (35)) can be computed as

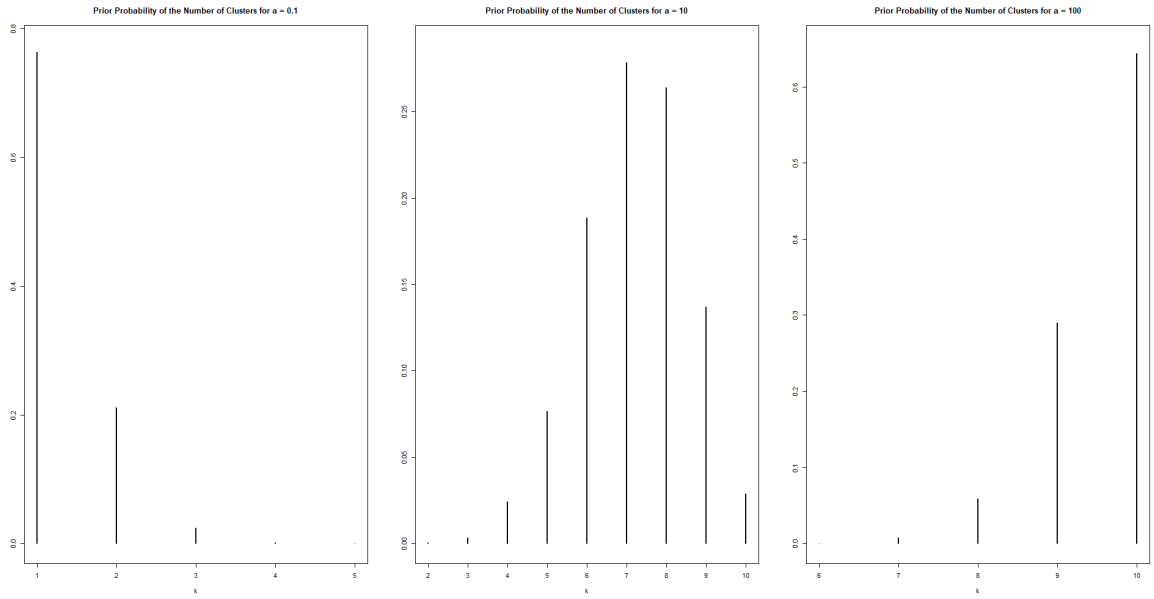
$$\mathbb{P}(k = l) = |s_1(n, l)| \, a^l \frac{\Gamma(a)}{\Gamma(a+n)}, l = 1, \ldots, n,$$

where $s_1(n, l)$ is the Stirling number of the first kind for n, l.

We now plot the prior density for k using a Monte Carlo method for different values of a.

```
x11()
par(mfrow=c(1,3))
```

```
for(a in c(0.1,10,100)){
    n = 10
    theta = vector(length=n)
    theta[1] = rnorm(1)
        for(j in 2:n){
        w0 = a/(j-1+a)
        w1 = rep(1/(j-1+a),j-1)
        index = sample(0:(j-1),size=1,prob=c(w0,w1))
        if(index==0){
            theta[j] = rnorm(1)
        }
        else{
            theta[j] = theta[index]
        }
    }
    k = vector(length=5000)
    for(i in 1:5000){
        theta[1] = rnorm(1)
        for(j in 2:n){
            w0 = a/(j-1+a)
            w1 = rep(1/(j-1+a),j-1)
            index = sample(0:(j-1),size=1,prob=c(w0,w1))
            if(index==0){
                theta[j] = rnorm(1)
            }
            else{
                theta[j] = theta[index]
            }
        }
        k[i] = length(unique(theta))
    }
    ymax = max(table(k)/5000)+0.01
    plot(table(k)/5000,ylim=c(0,ymax),ylab="")
    title(paste("Prior Probability of the Number of Clusters for", "a =",
        a, sep = " "))
}
```



In particular, observe that for a small the density of k is highly concentrated towards 1 *whereas for a large it is highly concentrated towards n.*

8.2 Dirichlet Process Mixture

8.2.1 The Dirichlet Process Mixture Model

The Dirichlet process generates distributions that are discrete with probability one, making it awkward for continuous density estimation. This limitation can be fixed by convolving its trajectories with some continuous kernel, or more generally, by using a Dirichlet process random measure as the mixing measure in a mixture over some simple parametric forms. Let Θ be a (typically finite dimensional) parameter space and, for each $\theta \in \Theta$, let $k(y, \theta)$ be a continuous probability density function. Given a probability distribution P defined on Θ, a mixture of $k(y, \theta)$ with respect to P has the following probability density function:

$$f(y) = \int_{\Theta} k(y, \theta) P(\mathrm{d}\theta)$$

Such mixture model together with a Dirichlet process prior on the mixing measure P gives the **Dirichlet process mixture model**:

$$\begin{aligned} Y_1, \ldots, Y_n | P &\overset{\text{iid}}{\sim} f(y) = \int_{\Theta} k(y, \theta) P(\mathrm{d}\theta) \\ P &\sim \mathcal{D}_{\alpha} \end{aligned} \tag{37}$$

where α is a finite measure on the parameter space Θ. In particular, some typical choices for the parametric kernel $k(y, \theta)$ are the following:

- $k(y, \theta) =$ density of $\mathcal{N}(\theta, 1)$.
- $k(y, \theta) =$ density of $\mathcal{N}\left(\mu, \sigma^2\right)$ with $\theta = \left(\mu, \sigma^2\right)$.
- $k(y, \theta)$ may be the density of the gamma or the Weibull distribution if the Y_i's are non-negative.

Remark. *If we express P via the stick-breaking construction ($P = \sum_{j=1}^{+\infty} V_j \delta_{\tau_j}$) then we have that $f(y) = \int_{\Theta} k(y, \theta) P(d\theta) = \sum_{j=1}^{+\infty} k(y, \tau_j) V_j$.*

An equivalent representation for the Dirichlet process mixture model is the following:

$$\begin{aligned} Y_i|\theta_i &\overset{\text{ind}}{\sim} k\left(\cdot,\theta_i\right), i=1,\ldots,n \\ \theta_1,\ldots,\theta_n|P &\overset{\text{iid}}{\sim} P \\ P &\sim \mathcal{D}_\alpha. \end{aligned} \tag{38}$$

Let us prove that (38) implies (37):

$$\begin{aligned} \mathcal{L}\left(Y_1=y_1,\ldots,Y_n=y_n|P\right) &= \int_{\Theta^n} \mathcal{L}\left(Y_1,\ldots,Y_n,d\theta_1,\ldots,d\theta_n|P\right) \\ &= \int_{\Theta^n} \mathcal{L}\left(Y_1,\ldots,Y_n|\theta_1,\ldots,\theta_n\right)\mathcal{L}\left(d\theta_1,\ldots,d\theta_n|P\right) \\ &= \int_{\Theta^n} \prod_{i=1}^{n} \mathcal{L}\left(Y_i|\theta_i\right)P\left(d\theta_i\right) \\ &= \int_{\Theta} k\left(y_1,\theta_1\right)P\left(d\theta_1\right)\cdot\ldots\cdot\int_{\Theta} k\left(y_n,\theta_n\right)P\left(d\theta_n\right) \end{aligned}$$

so that $Y_1,\ldots,Y_n|P \overset{\text{iid}}{\sim} f\left(y\right)=\int_\Theta k\left(y,\theta\right)P\left(\mathrm{d}\theta\right)$.

It is possible to prove that, under this hierarchical model (38) (or (37)), the posterior distribution of P is a mixture of Dirichlet processes:

$$P|Y_1,\ldots,Y_n \sim \int_{\Theta^n} \mathcal{D}_{\alpha+\sum_{i=1}^{n}\delta_{\theta_i}} H\left(d\theta_1,\ldots,d\theta_n|Y_1,\ldots,Y_n\right).$$

However, the analytic expression of the posterior $H\left(d\theta_1,\ldots,d\theta_n|Y_1,\ldots,Y_n\right)$ is not simple to be numerically evaluated because it contains the product of $n!$ factors. There are MCMC algorithms that can sample from the posterior of (38).
The Bayesian estimate of the density $f\left(y\right)$ is given by:

$$\mathbb{E}\left[f\left(y\right)|Y_1,\ldots,Y_n\right]=\mathbb{E}\left[\int_\Theta k\left(y,\theta\right)P\left(\mathrm{d}\theta\right)\middle|Y_1,\ldots,Y_n\right]=\int_\Theta k\left(y,\theta\right)\left(\mathbb{E}\left[P|Y_1,\ldots,Y_n\right]\left(\mathrm{d}\theta\right)\right),$$

where $\mathbb{E}\left[P|Y_1,\ldots,Y_n\right]$ is a probability on $\mathbb{R}$.

8.2.2 Clustering under the Dirichlet Process Mixture

An important implication of the previous formulation of the Dirichlet process mixture model (38) is that it induces a probability model on the clusters of the items $\{1,2,\ldots,n\}$, in the following sense.
The discrete nature of the Dirichlet process implies a positive probability for ties among the latent $\theta_i, i=1,\ldots,n$; see the second line in (38). Let $\theta_j^*, j=1,\ldots,k$ denote the $k\leq n$ unique values in $(\theta_1,\ldots,\theta_n)$, and let $S_l=\{i|\theta_j=\theta_l^*\}$ and $n_l=|S_l|$. Then $\rho_n=\{S_1,\ldots,S_k\}$ forms a partition of the set of experimental units $\{1,\ldots,n\}$ and, since the θ_i's are random, ρ_n itself is random. Note that ρ_n can assume only a finite number of values, though very large, that is the number of all partitions of $\{1,2,\ldots,n\}$.
In other words, the Dirichlet process mixture implies a prior on the random partition ρ_n of the experimental units, implied by the latent variables θ_i's. It is therefore possible to compute the posterior of ρ given the data $Y_1,..,Y_n$ and estimate the clustering structure of the data via a summary of such posterior distribution. In particular, a common choice is to choose the value of ρ minimizing a posteriori the expectation of some loss function L:

$$\widehat{\rho}=\arg\min_x \mathbb{E}\left[L\left(\rho,x\right)|Y_1,\ldots,Y_n\right]. \tag{39}$$

This is known as the **Bayesian nonparametric model-based approach to clustering**. Of course, the prior for ρ_n might be different that the one induced by the DP. However, in general, once that we assign the conditional distribution of datapoints, given ρ_n and cluster-specific parameters, we compute the posterior of ρ_n given data, and summarize it as in (39).

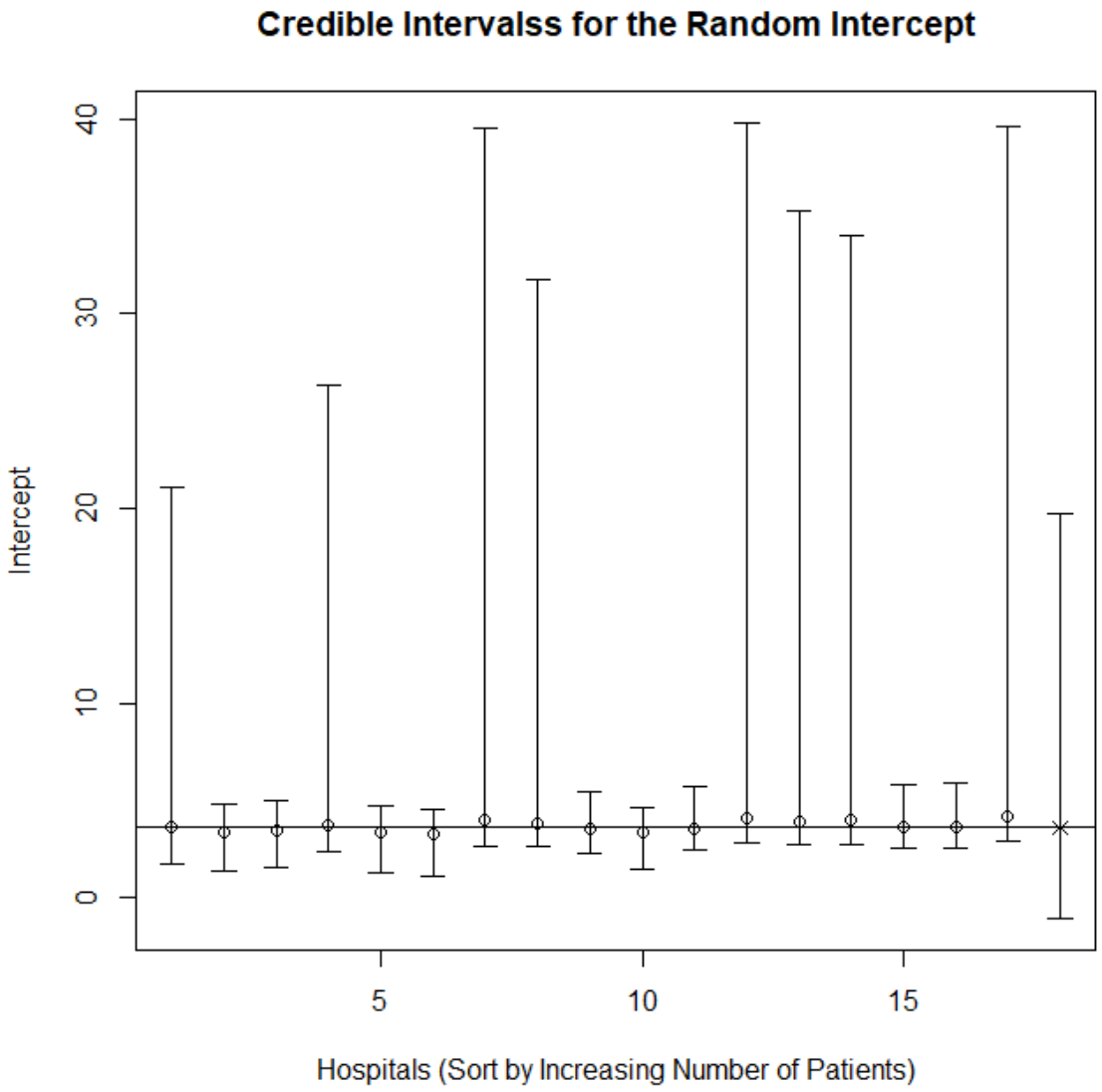
Example 8.4. *We now approach the task of performing cluster analysis by assuming a generalized linear mixed effect model in which the group specific random effects are sampled from a Dirichlet process. We consider a dataset regarding the mortality of infarction patients who went under angioplasty in $J = 17$ different hospitals. In particular, let Y_i be a binary random variable representing the status of the i-th patient after the surgery ($Y_i = 1$ if the i-th patient is alive and vice versa) and let $\boldsymbol{x}_i$ be a vector of covariates comprising the age of the patient (x_{i1}), the logarithm of the that elapsed before the surgery (x_{i2}), the infarction severity (x_{i3}) and the hospital to which the patient was admitted ($j\,[i]$). The model we consider is the following:*

$$
\begin{aligned}
Y_i|p_i &\overset{ind}{\sim} Be\left(p_i\right), \log\frac{p_i}{1-p_i} = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + b_{j[i]} \\
(\beta_1, \beta_2, \beta_3) &\perp\!\!\!\perp b_j \forall j \\
(\beta_1, \beta_2, \beta_3) &\sim \mathcal{N}\left(0, 100\mathbb{I}_3\right) \\
b_1, \ldots, b_J|P &\overset{iid}{\sim} P \\
P &\sim DP\left(\alpha, P_0\right)
\end{aligned}
$$

The partition of the hospital labels $\{1, \ldots, J\}$ is identified by the unique values of the b_j's.

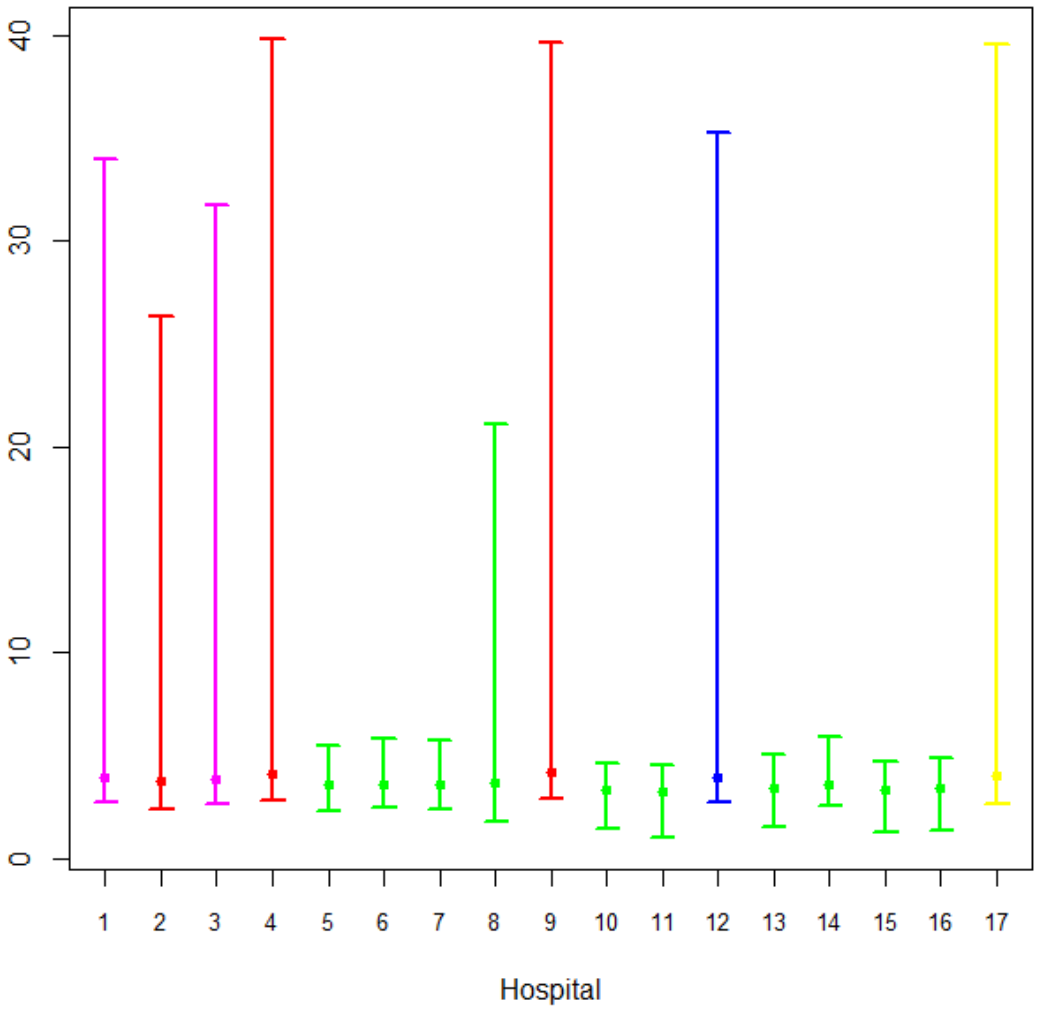
```
library(rjags)
library(plotrix)
library(coda)
input = read.table("data.txt",header=T)
n.data = dim(input)[1] # Number of patients
J = 17 # Number of hospitals
M = 40 # Truncation level for the stick-breaking construction
## Generate list of the data for JAGS
data = list(Y=input$vivo, AGE=input$eta,
            KILLIP=input$killip,
            LOGOB = input$logOB,
            CENTRO=input$centro,
            n=n.data, J=J, M=M)
## Define the initial state of the chain
r = rep(0.5,M)
theta = rep(0,M)
S = rep(1,J)
inits = list(beta=rep(0,3),a = 1,
            lambda.bb = 1,
            r=r,
            theta= theta,
            S = S,
            Snew = 1,
            .RNG.seed = 2,
            .RNG.name = 'base::Wichmann-Hill'
)
## Create the JAGS model
modelGLMM_DP = jags.model("model.bug",data=data,inits=inits,n.adapt=1000,
    n.chains=1)
update(modelGLMM_DP,19000) # Update for 19000 iterations without saving
```

```
variable.names = c("bb", "beta", "tau.bb", "newcentro", "alpha","K")
n.iter = 50000
thin = 10
outGLMM_DP = coda.samples(model=modelGLMM_DP,variable.names=variable.
    names,n.iter=n.iter,thin=thin)
data = as.matrix(outGLMM_DP)
data = data.frame(data)
attach(data)
## Credible Intervals for the Random Intercepts
# Sort the hospital by the number of patients
patients=rep(0,J)
for(i in 1:J){
patients[i]=length(which(input$centro==i))
}
sort_patients = sort(patients,index.return=T)
# Compute the quantiles for the random intercept
Q = matrix(nrow=J+1, ncol=3)
for(j in 1:J){
    Q[j,] = quantile(data[,2 + sort_patients$ix[j]],probs=c
        (0.025,0.5,0.975))
}
Q[J+1,] = quantile(data$newcentro,probs=c(0.025,0.5,0.975)) # This is the
     new hospital
colnames(Q) = c("2.5","median","97.5")
# Plot of the CIs (the last one is the new hospital)
pch = c(rep(21,J),4)
x11()
plotCI(x=seq(1,J+1), y=Q[,2], uiw=(Q[,3]-Q[,2]), liw=(Q[,2]-Q[,1]), pch=
    pch,
        scol=1 , xlab="Hospitals (Sort by Increasing Number of Patients)"
            ,
        ylab="Intercept", main="Credible Intervalss for the Random
            Intercept")
abline(h = mean(Q[,2]))
```




```
## Cluster Estimation with Binder Loss Function
label.mat = as.matrix(data[,3:19]) # Cluster labels
m = J
G = dim(data)[1]
pihat = matrix(0,ncol=m,nrow=m)
for(i in 1:G){
    ss = label.mat[i,]
    cij = outer(ss,ss,'==')
    pihat = pihat + cij
}
pihat = pihat/G
FF = vector("numeric")
K = 0.7
for(i in 1:G){
    ss = label.mat[i,]
    cij = outer(ss,ss,'==')
    binder = (pihat-K)*as.matrix(cij)
```

```
    binder = binder[upper.tri(binder)]
    FF[i] = sum(binder)
}
ind.bind = which.max(FF)[1]
ll.bind = label.mat[ind.bind,]
unici = unique(ll.bind) # Labels of the clusters
ncl = length(unici) # Number of clusters
for(i in 1:ncl){
    print(as.numeric(which(ll.bind==unici[i])))
}
# [1] 1 3
# [1] 2 4 9
# [1] 5 6 7 8 10 11 13 14 15 16
# [1] 12
# [1] 17
```



The corresponding JAGS code is given below.

```
model{
    # Precision Parameter
    a ~ dexp(1);
    alpha = a + 0.5; # Shifted Exp with support (0.5, +infinity) to avoid
        numerical issues

    # Stick-Breaking Construction
    ## Weights
    pp[1] = r[1];
    for (j in 2:M){
        pp[j] = r[j] * (1 - r[j - 1]) * pp[j - 1] / r[j - 1];
    }
    p.sum = sum(pp[]);

    ## Y_i's
    for (j in 1:M){
        theta[j] ~ dnorm(mu.bb,tau.bb); # Here a_0 is normal
        r[j] ~ dbeta(1, alpha);
        pi[j] = pp[j] / p.sum; # Renormalization of the weights
    }

    mu.bb = 0;
    tau.bb = pow(lambda.bb,-2);
    lambda.bb ~ dunif(0,50);
    for(i in 1:J){
        S[i] ~ dcat(pi[]); # Sampling from the mixture
        bb[i] = theta[S[i]];
        for (j in 1 : M){
            SC[i,j] = equals(j, S[i]);
        }
    }

    # Likelihood
    for(i in 1:n){
        logit(p[i]) = beta[1]*AGE[i] + beta[2]*LOGOB[i] + beta[3]*KILLIP[
            i] + bb[CENTRO[i]];
        Y[i] ~ dbern(p[i]);
    }
    # Prior for the Fixed Effects
    for(i in 1:3){
        mu[i] = 0;
        beta[i] ~ dnorm(mu[i], 0.001);
    }

    # New Random Hospital
    Snew ~ dcat(pi[]);
    newcentro = theta[Snew];

    # Clusters
    K = sum(cl[])
    for(j in 1:M){
        sumSC[j] = sum(SC[,j])
        cl[j] = step(sumSC[j] - 1)
    }
```

```
}
```

References

Albert, J. (2009). *Bayesian computation with R*. Springer.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Lee, D. (2022). CARBayes version 5.3: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors. *CRAN Vignette*.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Regazzini, E. (1996). Impostazione non parametrica di problemi d'inferenza statistica bayesiana.

Rockova, V., Lesaffre, E., Luime, J., and Löwenberg, B. (2012). Hierarchical bayesian formulations for selecting variables in regression models. *Statistics in medicine*, 31(11-12):1221–1237.

Rosner, G. L., Laud, P. W., and Johnson, W. O. (2021). *Bayesian thinking in biostatistics*. CRC Press.

Sahu, S. (2022). *Bayesian Modeling of Spatio-Temporal Data with R*. Chapman and Hall/CRC.

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728.

A Appendix: Review of Fundamentals of Probability

A.1 Notable Distributions

A.1.1 The gamma distribution

Definition A.1 (Gamma function)**.** *We define the* ***gamma function*** *as* $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}dx$*, for any* $\alpha > 0$*.*

Proposition A.1. *The following holds:*

- $\Gamma(1) = 1$ *and* $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$*.*
- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$*.*
- $\Gamma(n+1) = n!$ *for* $n \in \mathbb{N}$*.*

After the definition of the gamma function, we can introduce the gamma distribution.

Definition A.2 (gamma distribution)**.** *Let* X *be a real-valued random variable. We say that* X *is* ***gamma distributed with shape*** $\alpha > 0$ ***and rate*** 1 *if the distribution of* X *is absolutely continuous (with respect to the Lebesgue measure on* $\mathbb{R}$*) with density*

$$f_X(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-x}\mathbb{I}_{[0,+\infty)}(x).$$

In this case we write $X \sim gamma(\alpha, 1)$*.*

It is possible to extend the previous definition to any rate $\beta > 0$. To do that, let us first consider the following theorem.

Theorem A.1. *Let* X *be a random variable on* $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$ *and consider a map* $F : \left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right) \to \left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$*. If we define* $Y = F(X)$ *we have that* Y *has a density if the following conditions hold:*

1. *X has a density with respect to the Lebesgue measure.*
2. *F is invertible and has continuous partial derivatives.*

In this case the density of Y *is given by* $f_Y(y) = f_X\left(F^{-1}(y)\right) \cdot \left|\det JF^{-1}(y)\right| = f_X\left(F^{-1}(y)\right) \cdot \frac{1}{|\det JF(F^{-1}(y))|}$*.*

Remark. *In the simple case* $d = 1$ *we have* $f_Y(y) = f_X\left(F^{-1}(y)\right) \cdot \left|\frac{d}{dy}F^{-1}(y)\right|$*.*

Definition A.3. *Let* $X \sim gamma(\alpha, 1)$ *and let* $\beta > 0$*. We say that* $Y := \frac{X}{\beta}$ *is* ***gamma distributed with shape*** $\alpha > 0$ ***and rate*** β *and we write* $Y \sim gamma(\alpha, \beta)$*.*

Remark. *Starting from the previous theorem it is easy to check that if* $Y \sim gamma(\alpha, \beta)$ *then* Y *is an absolutely continuous r.v. with density*

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\mathrm{e}^{-\beta y}\mathbb{I}_{[0,+\infty)}(y).$$

Proposition A.2. *The following holds:*

- $gamma(1, \beta) \overset{d}{=} Exp(\beta)$*.*
- *If* $n \in \mathbb{N}$ *then* $gamma\left(\frac{n}{2}, \frac{1}{2}\right) \overset{d}{=} \chi^2(n)$*.*

- *If* $X \sim gamma(\alpha, \beta)$ *then* $Y = \frac{1}{X} \sim inv - gamma(\alpha, \beta)$ *and*

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha+1} e^{-\frac{\beta}{y}} \mathbb{I}_{(0,+\infty)}(y).$$

Moreover,

$$\mathbb{E}[Y] = \frac{\beta}{\alpha - 1} \text{ if } \alpha > 1 \quad \text{and } \operatorname{Var}(Y) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \text{ if } \alpha > 2.$$

Proposition A.3. *Let* $X \sim gamma(\alpha, \beta)$*. Then we have that:*

- *If* $\alpha > 1$*, the mode of* X *is* $\frac{\alpha-1}{\beta}$*.*
- $\mathbb{E}[X^n] = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\beta^n} = \frac{(\alpha+n-1)(\alpha+n-2)\cdot\ldots\cdot\alpha}{\beta^n}$*.*
- $\mathbb{E}[X] = \frac{\alpha}{\beta}$ *and* $\operatorname{Var}(X) = \frac{\alpha}{\beta^2}$*.*
- *If* $X_j \overset{ind}{\sim} gamma(\alpha_j, \beta), j = 1, \ldots, n$*, then* $\sum_{j=1}^n X_j \sim gamma\left(\sum_j^n \alpha_j, \beta\right)$ ***(additivity)****.*
- *If* $X_j \overset{ind}{\sim} gamma(\alpha_j, \beta), j = 1, 2$*, then* $X_1 + X_2$ *and* $\frac{X_j}{X_1+X_2}$ *are independent for both* $j = 1, 2$*.* ***(independence)****.*

A.1.2 The beta distribution

Definition A.4 (beta distribution)**.** *Let* X_1 *and* X_2 *be two random variables such that* $X_j \overset{ind}{\sim} gamma(\alpha_1, \beta)$*,* $j = 1, 2$*. We say that* $Y = \frac{X_1}{X_1+X_2}$ *is* ***beta distributed with parameters*** α_1 ***and*** α_2 *and we write* $Y \sim beta(\alpha_1, \alpha_2)$*.*

Proposition A.4. *If* $Y \sim beta(\alpha_1, \alpha_2)$ *then* Y *has density (wrt the Lebesgue measure on* $\mathbb{R}$*)*

$$f_Y(y) = \frac{1}{B(\alpha_j, \alpha_2)} y^{\alpha_1 - 1} (1-y)^{\alpha_2 - 1} \mathbb{I}_{(0,1)}(y)$$

where $B(\alpha_1, \alpha_2)$ *is the* ***beta function*** *and is defined as* $B(\alpha_1, \alpha_2) = \int_0^1 y^{\alpha_1-1}(1-y)^{\alpha_2-1} dy = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$*.*

Proposition A.5. *Let* $X \sim beta(\alpha_1, \alpha_2)$*. The following holds:*

- $\mathbb{E}[X^n] = \frac{B(\alpha_1+n, \alpha_2)}{B(\alpha_1, \alpha_2)}$ *for any* $n = 1, 2, \ldots$*.*
- $\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1+\alpha_2}$ *and* $\operatorname{Var}(X) = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2+1)(\alpha_1+\alpha_2)^2}$*.*
- $1 - X \sim beta(\alpha_2, \alpha_1)$*.*
- $beta(1, 1) \overset{d}{=} \mathcal{U}([0, 1])$*.*

A.1.3 The Dirichlet distribution

Definition A.5 (Dirichlet distribution)**.** *Let $X_1, \ldots, X_K$ be random variables such that $X_j \overset{ind}{\sim} gamma\,(\alpha_j, \beta)$, $j = 1, \ldots, K$. If we define $Y_i = \dfrac{X_i}{\sum_{j=1}^{n} X_j}$ for all i, then we say that $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$ is **Dirichlet distributed with parameters** $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{K-1}, \alpha_K)$ and we write $\boldsymbol{Y} \sim Dir\,(\alpha_1, \ldots, \alpha_K)$.*

Proposition A.6. *If $\boldsymbol{Y} \sim Dir\,(\alpha_1, \ldots, \alpha_K)$, then $(Y_1, \ldots, Y_{K-1})$ has a density, with respect to the Lebesgue measure on $\mathbb{R}^{K-1}$, given by*

$$f\,(y_1 \ldots, y_{K-1}) = \frac{1}{B\,(\boldsymbol{\alpha})} \prod_{j=1}^{K-1} y_j^{\alpha_j - 1} (1 - y_1 - y_2 - \cdots - y_{K-1})^{\alpha_K - 1} \mathbb{1}_{S_{K-1}}\,(y_1, \ldots, y_{K-1})$$

where $B\,(\boldsymbol{\alpha}) = \dfrac{\prod_{i=1}^{K} \Gamma\,(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$ and $S_{K-1} = \{(x_1, \ldots, x_{K-1}) \in \mathbb{R}^{K-1} : 0 \leq x_i \leq 1, i = 1, \ldots, K-1, \;\; 0 \leq x_1 + \cdots + x_{K-1} \leq 1\}$ is the $K-1$-dimensional simplex.

Remark. *Note that $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$ cannot have a density wrt the Lebesgue measure on $\mathbb{R}^K$, since the random vector $\boldsymbol{Y}$ is degenerate, i.e., $\sum_{i=1}^{K} Y_i = 1$ or, equivalently, $Y_K = 1 - Y_1 - Y_2 - \cdots - Y_{K-1}$.*

Proposition A.7. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_K) \sim Dir\,(\alpha_1, \ldots, \alpha_n)$. We have that:*

- *for all $j = 1, \ldots, K$, the marginal distribution of Y_j is such that $Y_j \sim beta\,(\alpha_j, \alpha_0 - \alpha_j)$ where $\alpha_0 = \sum_{i=1}^{n} \alpha_i$;*
- $\mathrm{Cov}\,(Y_i, Y_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)} < 0$ *for all $i \neq j$;*
- *if the random variables Y_i and Y_j are dropped out from the vector and replaced by their sum we have that $\boldsymbol{Y}' = (Y_1, \ldots, Y_i + Y_j, \ldots, Y_n) \sim Dir\,(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_n)$ (**aggregation**), i.e. $(Y_1 + Y_2, Y_3, \ldots, Y_K) \sim Dir(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$.*

A.1.4 The Multivariate Student's t distribution

Definition A.6. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^T$ be a random vector. We say that $\boldsymbol{Y}$ is t distributed with location $\boldsymbol{\mu}$, scale Σ and ν degrees of freedom if $\boldsymbol{Y}$ is an absolutely continuous vector with density*

$$f_{\boldsymbol{Y}}\,(\boldsymbol{y}) = \frac{\Gamma\left(\frac{q+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{(\det \Sigma)\,(\nu\pi)^q}} \left(1 + \frac{1}{\nu}\,(\boldsymbol{y} - \boldsymbol{\mu})^T\,\Sigma^{-1}\,(\boldsymbol{y} - \boldsymbol{\mu})\right)^{-\frac{q+\nu}{2}}.$$

In this case we write $\boldsymbol{Y} \sim t_q\,(\boldsymbol{\mu}, \Sigma, \nu)$.

Proposition A.8. *If $\boldsymbol{Y} \sim t_q\,(\boldsymbol{\mu}, \Sigma, \nu)$ we have that:*

- *If $\nu > 1$ then $\mathbb{E}\,[\boldsymbol{Y}] = \boldsymbol{\mu}$.*
- *If $\nu > 2$ then $Var\,(\boldsymbol{Y}) = \frac{\nu}{\nu-2}\Sigma$.*

A.1.5 The Wishart and inverse Wishart distributions

Definition A.7. *A matrix M is positive definite if $\boldsymbol{z}^T M \boldsymbol{z} > 0$ for all $\boldsymbol{z} \neq \mathbf{0}$.*

Definition A.8. *A $p \times p$ matrix X, symmetric and positive definite, has Wishart density with parameter (M, ν), where $\nu \geq p$ and M symmetric and positive definite $p \times p$ matrix:*

$$f(X|M,\nu) = \frac{1}{2^{\frac{\nu p}{2}} \Gamma_p(\frac{\nu}{2}) |M|^{\nu/2}} |X|^{\frac{\nu-p-1}{2}} \exp\left\{-tr\left(\frac{M^{-1}X}{2}\right)\right\},$$

where $\Gamma_p(\frac{\nu}{2}) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(\frac{\nu+1-j}{2})$ and $tr(A) = \sum_j a_{jj}$ (sum of the diagonal elements of the matrix).
In this case we write $X \sim Wishart(M, \nu)$.

An alternative definition of the Wishart distribution considers $\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, \Sigma)$, $i = 1, \ldots, n$ and Σ positive definite. Then,

$$X = \sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i^T \sim Wishart(\Sigma, n)$$

Example A.1. *When $p = 2$*

$$X = \begin{bmatrix} \sum_{i=1}^n Y_{i1}^2 & \sum_{i=1}^n Y_{i1}Y_{i2} \\ \sum_{i=1}^n Y_{i1}Y_{i2} & \sum_{i=1}^n Y_{i2}^2 \end{bmatrix} \sim Wishart(\Sigma, n).$$

Properties of the Wishart distribution:

- If $X \sim Wishart(M, \nu)$, then

$$\begin{aligned} \mathbb{E}(X|M,\nu) &= \nu M \\ \text{Var}(X_{ij}|M,\nu) &= \nu(m_{ij}^2 + m_{ii}m_{jj}), \quad \text{where } X = [X_{ij}]_{ij},\ M = [M_{ij}]_{ij} \end{aligned}$$

- It is a multivariate version of the gamma distribution; in fact, if $p = 1$ and $M = 1$, then $X \sim \chi^2(\nu) = gamma(\nu/2, 1/2)$.

- The Wishart density is the conjugate prior for precision matrix Σ^{-1} under the multivariate normal likelihood:

$$\boldsymbol{Y}_i|\Sigma \overset{\text{iid}}{\sim} N_p(\boldsymbol{\mu}_0, \Sigma), \quad \Sigma^{-1} \sim Wishart(M, \nu) \text{ then } \pi(\Sigma^{-1}|\boldsymbol{y}) \sim Wishart(\cdot, \cdot).$$

Definition A.9. *A $p \times p$ matrix W, symmetric and positive definite, has inverse-Wishart density with hyperparameters (M^{-1}, ν) if $W^{-1} \sim Wishart(M, \nu)$.*
In this case we write $W \sim inverse - Wishart(M^{-1}, \nu)$.

If W is a $p \times p$ matrix and $W \sim inverse - Wishart(M^{-1}, \nu)$ then

$$\mathbb{E}(W) = \frac{1}{\nu - p - 1} M^{-1}, \quad \nu > p + 1$$

A.2 Conditional Probability

We **informally** revise definitions and properties of conditional distributions.

Definition A.10 (Conditional probability)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $A, B \in \mathcal{F}$ *be two events such that* $\mathbb{P}(B) > 0$*. We define* ***conditional probability of*** **A** ***given*** **B** *the quantity* $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$*.*

We will now recall how one can define the conditional distribution $\mathcal{L}(X|Y)$ of a random variable X given another random variable Y (both defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$). We recall first the definition of conditional probability and distribution, given a sub-σ-field $\mathcal{G}$, that, in the rest of the course, will be the σ-field generated by another random variable Y. We focus on random variables, but we could consider more general random elements with values in a Polish space.

Definition A.11. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space; let* $\mathcal{G}$ *be a sub-*σ*-filed of* $\mathcal{F}$ *and let and consider an event* $A \in \mathcal{F}$*. We define the conditional probability of* A *given* $\mathcal{G}$*,* $\mathbb{P}(A|\mathcal{G})$*, as the random variable that satisfies the following conditions:*

1. $\omega \mapsto \mathbb{P}(A|\mathcal{G})(\omega)$ *is* $\mathcal{G}$*-measurable and* $\mathcal{G}$*-integrable.*
2. *it must satisfy the following integral equation:*

$$\mathbb{P}(A \cap G) = \int_G \mathbb{P}(A|\mathcal{G})(\omega)\, \mathbb{P}(d\omega) \text{ for all } G \in \mathcal{G}.$$

The definition can be extended to define the conditional distribution.

Definition A.12. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* X *be a random variable defined on the same space; let* $\mathcal{G}$ *be a sub-*σ*-filed of* $\mathcal{F}$*. We define the conditional distribution of* X *given* $\mathcal{G}$*,* $\mu(H, \omega)$ *as a function* $\mu : \mathcal{B}(\mathbb{R}) \times \Omega \to \mathbb{R}$ *such that:*

1. $H \mapsto \mu(H, \omega)$ *is a probability on* $\mathbb{R}, \mathcal{B}(\mathbb{R})$ *for all* $\omega \in \Omega$*;*
2. $\omega \mapsto \mu(H, \omega)$ *is a* ***version*** *of the conditional probability* $\mathbb{P}(X \in H|\mathcal{G})(\omega))$*, i.e.,*

$$\mathbb{P}(\{X \in H\} \cap G) = \int_G \mu(H, \omega) P(d\omega) \text{ for all } G \in \mathcal{G}.$$

We use notation $\mathbb{P}(X \in H|\mathcal{G})$ *to denote the r.v.* $\mu(H, \omega)$*.*

We now consider two r.v.'s X and Y, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and define the conditional distribution of X given Y, denoted by as $\mathcal{L}(X|Y)$ in a special case. We apply the previous definition when $\mathcal{G} = \sigma(Y)$, i.e., when $\mathcal{G}$ is the σ-field generated by the r.v. Y. We assume that the joint distribution of (X, Y) has a density $f(x, y)$, w.r.t. some reference measure on $\mathbb{R} \times \mathbb{R}$ (for instance, the Lebesgue measure, that is when (X, Y) is an absolutely continuous random vector).

Definition A.13. *If* $(X, Y) \sim f(x, y)$ *where* $f(x, y)$ *is a density function on* $\mathbb{R} \times \mathbb{R}$*, then the conditional distribution* $\mathcal{L}(X|Y)$ *has density*

$$f_{X|Y}(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & \text{for } y \text{ such that } f_Y(y) > 0 \\ const & \text{otherwise} \end{cases}. \tag{40}$$

where $f_Y(y) = \int_{\mathbb{R}} f(x, y)\, dx$ *is the marginal density of* X*.*

Remark. *Note that* (40) *yields a method to assign the joint distribution of* (X, Y) *as*

$$f(x, y) = f_{X|Y}(x|y) \times f_Y(y).$$

In this case, the **conditional expectation** and **conditional variance** of X, given Y are easily introduced as

$$\mathbb{E}[X|Y] = \int_{\mathbb{R}} x f_{X|Y}(x|y)\,\mathrm{dx},$$

and

$$\mathrm{Var}(X|Y) = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2 \middle| Y\right].$$

We now informally revise useful properties of the conditional expectation.

Proposition A.9. *The following properties hold:*

1. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$.
2. $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\mathrm{Var}(X|Y)]$.
3. *If* $c \in \mathbb{R}$ *then* $\mathbb{E}[c|Y] = c$.
4. $\mathbb{E}[aX_1 + bX_2|Y] = a\mathbb{E}[X_1|Y] + b\mathbb{E}[X_2|Y]$.
5. *If* $X \geq 0$ *almost surely (a.s.), then* $\mathbb{E}[X|Y] \geq 0$ *a.s.*.
6. *If* $X_1 \geq X_2$ *a.s., then* $\mathbb{E}[X_1|Y] \geq \mathbb{E}[X_2|Y]$ *a.s.*.
7. $\mathbb{E}[X|X] = X$ *a.s.*.
8. $\mathbb{E}[Xg(Y)|Y] = g(Y)\,\mathbb{E}[X|Y]$ *where* $g(Y)$ *is a deterministic function of* Y.

For more details see Billingsley (2017).

B Appendix: Well-known Bayesian Models

B.1 The Bernoulli-Beta Model

Assume that $Y_1, \ldots, Y_n | \theta \overset{\text{iid}}{\sim} \text{Be}(\theta)$ and that a priori $\theta \sim beta(\alpha, \beta)$ with $\alpha, \beta > 0$. It is straightforward to compute the conditional density of data, given θ as

$$f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum_1^n y_i} \times (1-\theta)^{n-\sum_1^n y_i}.$$

We compute the posterior via the Bayes' theorem (2):

$$\begin{aligned}\pi(\theta|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\theta)\,\pi(\theta)}{\int_{\mathbb{R}} f(\boldsymbol{y}|\theta)\,\pi(\theta)\,\mathrm{d}\theta} = \frac{\theta^{\sum_1^n y_i}(1-\theta)^{n-\sum_1^n y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{(0,1)}(\theta)}{\int_0^1 \theta^{\sum_1^n y_i}(1-\theta)^{n-\sum_1^n y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta} \\ &= \frac{\theta^{a+\sum_{i=1}^n y_i - 1}(1-\theta)^{n-\sum_{i=1}^n y_i+\beta-1}(\theta)}{\frac{\Gamma\left(\alpha+\sum_{i=1}^n y_i\right)\Gamma\left(\beta+n-\sum_{i=1}^n y_i\right)}{\Gamma(\alpha+\beta+n)}}\mathbb{1}_{(0,1)}(\theta)\end{aligned}$$

It is clear that

$$\theta|\boldsymbol{y} \sim beta\left(\alpha + \sum_{i=1}^n y_i, \beta + n - \sum_{i=1}^n y_i\right),$$

that is the posterior is still a beta distribution. This is a first example of **conjugate Bayesian model**, or equivalently, we say that **the prior is conjugate to the likelihood**. Observe that

$$\mathbb{E}_\pi[\theta|\boldsymbol{y}] = \frac{\alpha + \sum_{i=1}^n y_i}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\frac{1}{n}\sum_{i=1}^n y_i = w_n \mathbb{E}_\pi[\theta] + (1-w_n)\,\overline{y}$$

where $w_n := \frac{\alpha}{\alpha+\beta+n}$. Hence, the Bayesian point estimate of θ is given by a convex linear combination of the prior guess $\mathbb{E}_\pi[\theta]$ and the frequentist point estimate $\overline{y}$.
Moreover, we have that $w_n \xrightarrow{n\to+\infty} 0$ so that the Bayesian estimate tends to the frequentist estimate for n large. All in all, it is possible to show that $\text{Var}(\theta|\boldsymbol{y}) = \mathbb{E}_\pi[\theta|\boldsymbol{y}](1-\mathbb{E}_\pi[\theta|\boldsymbol{y}])\frac{1}{\alpha+\beta+1+n}$ so that $\text{Var}(\theta|\boldsymbol{y}) \xrightarrow{n\to+\infty} 0$.

B.2 The Normal-Normal Model

Assume that $Y_1, \ldots, Y_n | \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$, while σ_0^2 is known, and that a priori $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$. In this case, the conditional distribution of data $\boldsymbol{Y}$, given μ, is:

$$f(\boldsymbol{y}|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \mathrm{e}^{-\frac{(y_i-\mu)^2}{2\sigma_0^2}} = \frac{1}{(2\pi\sigma_0^2)^{\frac{n}{2}}} \mathrm{e}^{-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma_0^2}} = \frac{1}{(2\pi\sigma_0^2)^{\frac{n}{2}}} \mathrm{e}^{-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n y_i^2 - n\overline{y}^2\right)} \mathrm{e}^{-\frac{n}{2\sigma_0^2}(\mu-\overline{y})^2}$$

The posterior is proportional to the conditional density of $\boldsymbol{Y}$ given μ times the prior density. We have:

$$\pi(\mu|\boldsymbol{y}) \propto \frac{1}{(2\pi\sigma_0^2)^{\frac{n}{2}}} \mathrm{e}^{-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n y_i^2 - n\overline{y}^2\right)} \mathrm{e}^{-\frac{n}{2\sigma_0^2}(\mu-\overline{y})^2} \frac{1}{\sqrt{2\pi\tau_0^2}} \mathrm{e}^{-\frac{(\mu-\mu_0)^2}{2\tau_0^2}}$$

Let us now focus only on the terms depending on μ at the exponent. We have that:

$$\begin{aligned}-\frac{1}{2}\left[\frac{n}{\sigma_0^2}(\mu-\overline{y})^2 + \frac{1}{\tau_0^2}(\mu-\mu_0)^2\right] &= -\frac{1}{2}\left[\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau_0^2}\right)(\mu-\mu_n)^2 + \frac{\frac{n}{\sigma_0^2\tau_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\tau_0^2}}(\overline{y}-\mu_n)^2\right] \\ &= -\frac{1}{2}\left[\frac{(\mu-\mu_n)^2}{\tau_n^2} + \frac{\frac{n}{\sigma_0^2\tau_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\tau_0^2}}(\overline{y}-\mu_n)^2\right]\end{aligned}$$

where we have defined $\mu_n := \frac{n\tau_0^2\overline{y}+\sigma_0^2\mu_0}{n\tau_0^2+\sigma_0^2}$ and $\tau_n^2 := \frac{\sigma_0^2\tau_0^2}{n\tau_0^2+\sigma_0^2}$.

Remark. *We have made use of the well-known equality*

$$d_1 (z - c_1)^2 + d_2 (z - c_2)^2 = (d_1 + d_2)(z - c)^2 + \frac{d_1 d_2}{d_1 + d_2}(c_1 - c_2)^2$$

where $c := \frac{d_1 c_1 + d_2 c_2}{d_1 + d_2}$.

Hence, beyond multiplicative factors that do not depend on μ, the posterior is proportional to $\pi(\mu|\boldsymbol{y}) \propto \mathrm{e}^{-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\tau_n^2}}$, so that the posterior of μ is still a Gaussian density, i.e.,

$$\mu|\boldsymbol{y} \sim \mathcal{N}(\mu_n, \tau_n^2).$$

Note that

$$\mu_n := \mathbb{E}(\mu|\mathbf{y}) = \frac{\sigma_0^2}{n\tau_0^2 + \sigma_0^2}\mu_0 + \frac{n\tau_0^2}{n\tau_0^2 + \sigma_0^2}\overline{y} = w_n\mu_0 + (1 - w_n)\bar{y}$$
$$\tau_n^2 = \frac{\sigma_0^2\tau_0^2}{n\tau_0^2 + \sigma_0^2} = \frac{\sigma_0^2}{n\tau_0^2 + \sigma_0^2}\tau_0^2 < \tau_0^2$$

Remark. *The posterior mean is a convex linear combination of the prior mean* μ_0 *and the empirical mean* $\bar{y}$*. The posterior variance is smaller that the prior variance. If the sample size* n *is large, then the Bayesian estimate* μ_n *is close to* $\bar{y}$*, and the posterior variance around this estimate is small as well.*

Similarly, we can show that the posterior predictive distribution is given by $Y_{n+1}|Y_1, \ldots, Y_n \sim \mathcal{N}(\mu_n, \sigma_0^2 + \tau_n^2)$.