Università degli studi di Milano–Bicocca

Scuola di Economia e Statistica

Corso di Laurea in

Scienze Statistiche ed Economiche

An introduction to Bayesian mixture models and applications

Relatore: Dott. Federico Camerlenghi

Correlatore: Dott. Tommaso Rigon

Tesi di laurea di:

Anna Petranzan

Matricola N. 858541

Anno Accademico 2021/2022

Contents

Introduction

There is a substantial difference between the classical and the Bayesian approaches to probability. The first considers probability as an objective value, where parameters are fixed and unknown, and inferential procedures are based on repeated sampling under the same conditions. The latter is much more subjective. As a matter of fact, in the Bayesian approach parameters are considered random variables, and a prior distribution is assumed based on them, without considering the data. Only afterwards, when the data is observed, the assumptions made are updated to obtain the posterior distribution.

This thesis provides an introduction to the Bayesian approach, together with focusing on the density estimation for the mixture models using the Gibbs sampling, a particular technique which is well suited for the models of our interest.

The thesis consists of four chapters. In the first chapter some key tools of Bayesian statistics are explained in detail. The topics presented are very specific but, even though not all of them will be used in the rest of the paper, they are fundamental to understand the Bayesian approach to statistical inference.

The second chapter offers a general overview of already known sampling algorithms such as the Monte Carlo method and Monte Carlo Markov Chain. Finally, the Gibbs sampling method is introduced.

The third chapter focuses on mixture models. These are useful for density estimation for populations composed by different sub-populations, each having a different density distribution. Some of the difficulties that come with these specific models are taken into consideration here, such as choosing the number of components and the issues with label exchangeability when identifying sub-populations. At the end, the Gibbs sampling method is detailed in the framework of mixture models. These arguments are, in fact, necessary for understanding the implementation of the algorithm and the examples provided in the last chapter.

Chapter four is the core of this thesis, containing the description of the construction of the algorithm for the implementation in the R software. Brief descriptions of the implemented code are provided through the use of a theoretical example. In addition, there are several examples that include both basic datasets already directly available from R in different libraries, like Old Faithful Geyser Data and Galaxy Data, and a dataset of real data: Concrete Compressive Data. In this way, this study demonstrates and verifies the functioning of the implemented algorithm on data of varying complexity.

Chapter 1

Introduction to Bayesian inference

Within this chapter the definitions underlying Bayesian statistics will be discussed. This section will provide the statement of the Bayes' rule, as well as an explanation of the meaning of random variables (in both the discrete and discontinuous cases), the way independence is defined, and how such variables are distributed. We will explain how to estimate variables and the concept of exchangeability related to the concept of conjugacy.
Finally, the normal model that is the main character of this paper will be explored. The main reference is Hoff (2009).

Bayesian inference is based on Bayes' rule, which shows us how our beliefs should change once we observe a sample y from the sample space $\mathscr{Y}$.
The purpose of Bayesian inference is to use the data to quantify the reduction in uncertainty about population parameters.
It is important to note that Bayes' rule does not tell us what our beliefs on the parameter should be after observing evidence; rather, it informs us how they should change.

1.1 Bayes' rule

We denote the *parameter space* by Θ which is the collection of potential parameter values, as specified in the first chapter of Hoff (2009). The Bayesian method begins with a numerical expression of joint beliefs about y and θ, i.e. the sample and the parameter respectively, in terms of probability distributions over $\mathscr{Y}$ and

Θ. The *prior distribution* $p(\theta)$ expresses our view that θ represents real population features for each numerical value θ in Θ. The *sampling model* is represented by the conditional distribution of y given θ, $p(y|\theta)$, which expresses our hypothesis that y would be the outcome of our study if we knew θ was true for each $\theta \in \Theta$ and y in $\mathscr{Y}$.

We update our assumptions about θ once we observed the data y and we get the *posterior distribution* $p(\theta|y)$, which expresses our opinion that θ is the real value, given dataset y, as θ varies in Θ. Using the *Bayes' rule*, the posterior distribution is derived from the prior distribution and sampling model. In fact, the *Bayes' rule* is the tool to update the prior and get the posterior:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})\, d\tilde{\theta}} \,.$$

Bayes' rule is the best approach for updating beliefs about θ given new knowledge, according to Cox (1946), Cox (1961) and Savage (1954), Savage (1972). Their findings provide a solid theoretical foundation for using Bayes' rule as a quantitative learning approach. However, it might be difficult to accurately mathematically define our prior beliefs in practical data analysis settings, hence $p(\theta)$ is frequently used on an ad hoc basis or for computational efficiency.

1.2 Starting definitions

First of all, we introduce some definitions, in particular which of both continuous and discrete random variables in the Bayesian context. These are important notions for describing the sample we are going to consider and on which we will make inference. For such a sample, the hypothesis of exchangeability will be supposed, a fundamental assumption for the concept of conjugacy, both of which are explained here.

We refer to Chapter 2 of the Hoff (2009) for additional details.

1.2.1 Random variables

Definition 1.1 (Random variables)**.** Given $(\Omega, \mathcal{B}), (\mathbb{R}, \mathcal{B}_1)$ measurable spaces, where $\mathcal{B}_1$ is the σ-algebra over $\mathbb{R}$, a *random variable* is an application $Y : \Omega \to \mathbb{R}$

such that:

$$\forall B_1 \in \mathcal{B}_1 : Y^{-1}(B_1) \in \mathcal{B} \, .$$

In general, a random variable is an application from an underlying probability space to a set of interest, which can be the sample space $\mathcal{Y}$ or the parameter space Θ.

Definition 1.2 (Probability measure)**.** A random variable induces on measurable space $(\mathbb{R}, \mathcal{B}_1)$ a new probability measure, said distribution of Y, which we denote with $\mathbb{P}_Y$.

$$\mathbb{P}_Y : \mathcal{B}_1 \to \mathbb{R} \text{ such that } \forall B_1 \in \mathcal{B}_1 \; \mathbb{P}_Y(B_1) = \mathbb{P}(Y^{-1}(B_1)).$$

From now on, we will denote all probability distributions with $\mathbb{P}$, where the related variable will be clear from the context.

Discrete random variables

Let Y be a random variable and let $\mathcal{Y}$ be the set of all possible values of Y.
This variable is discrete if the set of possible outcomes is countable, meaning that $\mathcal{Y}$ can be expressed as $\mathcal{Y} = \{y_1, y_2, \dots\}$.

The set $\{Y = y\}$ contains all the possible outcomes of an experiment which yield the values y for the variable Y. The notation for $\mathbb{P}(Y = y)$ is shorted to $p(y)$ for each $y \in \mathcal{Y}$. The *probability mass function* (pmf) of Y is a function of y that has the following properties:

- $0 \leqslant p(y) \leqslant 1 \; \forall y \in \mathcal{Y}$;
- $\sum_{y \in \mathcal{Y}} p(y) = 1$.

The pmf may be used to generate general probability assertions regarding Y. For example, supposed A a subset in $\mathcal{Y}$, $\mathbb{P}(Y \in A) = \sum_{y \in A} p(y)$. If A and B are disjoint subsets of $\mathcal{Y}$, then

$$\mathbb{P}(Y \in A, Y \in B) \equiv \mathbb{P}(Y \in A \cup B) = \mathbb{P}(Y \in A) + \mathbb{P}(Y \in B) = \sum_{y \in A} p(y) + \sum_{y \in B} p(y).$$

Continuous random variables

Assume the sample space $\mathscr{Y}$ is equal to $\mathbb{R}$. As a result, event probabilities cannot be defined in terms of a pmf $p(y)$, but rather in terms of a *cumulative distribution function* (cdf):

$$F_Y(y) = \mathbb{P}(Y \leqslant y).$$

Note that $\lim_{y \to +\infty} F_Y(y) = 1$, $\lim_{y \to -\infty} F_Y(y) = 0$, and $F_Y(b) \leqslant F_Y(a)$ if $b < a$. Probabilities of various events can be derived from the cdf:

- $\mathbb{P}(Y > a) = 1 - F_Y(a)$;
- $\mathbb{P}(a < Y \leqslant b) = F_Y(b) - F_Y(a)$.

Definition 1.3 (Continuous random variable)**.** A random variable Y is called an absolutely continuous random variable if there exists a non-negative function p of $\mathbb{R}$ such that:

$$F_Y(Y \leqslant y) = \int_{-\infty}^{y} p(y)\, dy,\ \forall\, y \in \mathbb{R}.$$

This function is called the *probability density function* of Y and the following properties hold:

- $0 \leqslant p(y)\, \forall\, y \in \mathscr{Y}$;
- $\int_{\mathbb{R}} p(y)\, dy = 1$.

They are similar to those of a pdf for a discrete random variable. In addition, as in the discrete case, probability statements about Y can be derived from the pdf: $\mathbb{P}(Y \in A) = \int_A p(y)\, dy$, and if A and B are disjoint subsets of $\mathscr{Y}$, then

$$\begin{aligned}\mathbb{P}(Y \in A, Y \in B) \equiv \mathbb{P}(Y \in A \cup B) &= \mathbb{P}(Y \in A) + \mathbb{P}(Y \in B) \\ &= \int_A p(y)\, dy + \int_B p(y)\, dy.\end{aligned}$$

1.2.2 Joint distributions

In the Bayesian context, it is constantly used to work with several random variables simultaneously. In the following section, we will define how two random variables are jointly distributed.

Discrete joint distribution

Let $\mathscr{Y}_1$, $\mathscr{Y}_2$ be two countable sample spaces and Y_1, Y_2 be two random variables, taking values in $\mathscr{Y}_1$, $\mathscr{Y}_2$ respectively. Joint beliefs about Y_1 and Y_2 can be represented with probabilities. The *joint pdf* or *joint density* of Y_1 and Y_2 is defined as

$$p_{(Y_1,Y_2)}(y_1,y_2) = \mathbb{P}(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \quad \text{for } y_1 \in \mathscr{Y}_1,\ y_2 \in \mathscr{Y}_2.$$

The *marginal density* of Y_1 can be computed from the joint density:

$$\begin{aligned} p_{Y_1}(y_1) &\equiv \mathbb{P}(Y_1 = y_1) \\ &= \sum_{y_2 \in \mathscr{Y}_2} \mathbb{P}(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \\ &= \sum_{y_2 \in \mathscr{Y}_2} p_{(Y_1,Y_2)}(y_1,y_2). \end{aligned}$$

The *conditional density* of Y_1 can be computed from the joint density:

$$\begin{aligned} p_{Y_2|Y_1}(y_2|y_1) &= \frac{\mathbb{P}(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{\mathbb{P}(Y_1 = y_1)} \\ &= \frac{p_{Y_1Y_2}(y_1,y_2)}{p_{Y_1}(y_1)}. \end{aligned}$$

Continuous joint distributions

If Y_1 and Y_2 are absolutely continuous and given a continuous joint cdf $F_{(Y_1,Y_2)}(a,b) \equiv \mathbb{P}(\{Y_1 \leqslant a\} \cap \{Y_2 \leqslant b\})$, there is a function $p_{(Y_1,Y_2)}$ such that:

$$F_{(Y_1,Y_2)}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} p_{(Y_1,Y_2)}(y_1,y_2)\, dy_2 dy_1.$$

The function $p_{(Y_1,Y_2)}$ is the joint density of Y_1 and Y_2 and it holds that:

- $p_{Y_1}(y_1) = \int_{\mathbb{R}} p_{(Y_1,Y_2)}(y_1,y_2)\, dy_2$;

- $p_{Y_2|Y_1}(y_2|y_1) = \frac{p_{(Y_1,Y_2)}(y_1,y_2)}{p_{Y_1}(y_1)}$.

Mixed continuous and discrete variables

Let Y_1 be discrete and Y_2 be continuous. Define a marginal density p_{Y_1} from our beliefs $\mathbb{P}(Y_1 = y_1)$ and a conditional density $p_{Y_2|Y_1}(y_2|y_1)$ from $\mathbb{P}(Y_2 \leqslant y_2|Y_1 = y_1) \equiv F_{Y_2|Y_1}(y_2|y_1)$ as above. The joint density of Y_1 and Y_2 is then:

$$p_{(Y_1,Y_2)}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2|Y_1}(y_2|y_1).$$

1.2.3 Independent random variables

Assume that $Y_1, \ldots, Y_n$ are random variables and that θ is a parameter that describes the conditions under which the random variables are formed. It is important to remember that in the Bayesian context θ is a realisation of a random variable. The random variables $Y_1, \ldots, Y_n$ are *conditionally independent* given θ if for every collection of sets $\{A_1, \ldots, A_n\}$ we have

$$\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n|\theta) = \mathbb{P}(Y_1 \in A_1|\theta)\mathbb{P}(Y_2 \in A_2|\theta) \cdots \mathbb{P}(Y_n \in A_n|\theta). \quad (1.1)$$

The Equation (1.1) is based on the notion of independent events, where each $\{Y_j \in A_j\}$ represents an event.

Definition 1.4 (Independence of events)**.** Two events F and G are conditionally independent given H if $\mathbb{P}(F \cap G|H) = \mathbb{P}(F|H)\mathbb{P}(G|H)$.

From Equation (1.1) if independence holds, then

$$\mathbb{P}(Y_i \in A_i|\theta, Y_j \in A_j) = \mathbb{P}(Y_i \in A_i|\theta).$$

Knowing Y_j, according to this understanding, provides no extra knowledge about Y_i beyond what θ provides. The product of marginal densities gives the joint density under independence:

$$p(y_1, \ldots, y_n|\theta) = p_{Y_1}(y_1|\theta)p_{Y_2}(y_2|\theta) \cdots p_{Y_n}(y_n|\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i|\theta).$$

If the marginal densities are all equal to some common density, and it results that:

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta),$$

the $Y_1, \ldots, Y_n$ are *conditionally independent and identically distributed* (i.i.d).

1.3 Bayes' rule and parameter estimation

In the Bayesian context, parameter estimation allows us to understand how much we need to update our beliefs in accordance with the observed data. In this section, the estimation procedure will be explained.

Let consider the discrete random sample Y and the parameter θ. The joint distribution $p(y, \theta)$, which captures our opinions about θ and the survey outcome Y, is required to calculate the posterior distribution $p(\theta|y)$. It is frequently made up of:

- $p(\theta)$, beliefs about θ;
- $p(y|\theta)$, beliefs about Y for each value of θ.

Having observed $\{Y = y\}$, it is needed to update beliefs about θ:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}. \tag{1.2}$$

The conditional density in the Equation (1.2) is called the *posterior density* of θ. It results that:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{1.3}$$

The constant of proportionality in the Equation (1.3) is $1/p(y)$ which could be computed from:

$$p(y) = \int_{\Theta} p(y, \theta)\, d\theta = \int_{\Theta} p(\theta)p(y|\theta)\, d\theta.$$

It follows that:

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

1.4 Exchangeability

An important property that we will assume to be valid throughout the rest of the discussion, is that of *exchangeability*. It is very useful as it guarantees that for random variables having this property, the joint distribution does not change, whatever sequence they are taken with.

Definition 1.5. Let $p(y_1, \ldots, y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If $p(y_1, \ldots, y_n) = p(y_{\pi(1)}, \ldots, y_{\pi(n)})$ for all permutations π and $n \geqslant 1$, then $Y_1, \ldots, Y_n$ are exchangeable.

Definition (1.5) means that if $Y_1, \ldots, Y_n$ are exchangeable, labels convey no information about the outcomes.

Lemma 1.1. *If* $\theta \sim p(\theta)$ *and* $Y_1, \ldots, Y_n$ *are conditionally i.i.d given* θ*, then marginally (unconditionally on* θ*),* $Y_1, \ldots, Y_n$ *are exchangeable.*

1.4.1 de Finetti's theorem

From the exchangeability property, the de Finetti theorem states that exchangeable observations are i.i.d conditional on some latent variable.

Theorem 1.1. *Let* $Y_i \in \mathscr{Y}$ *for all* $i \in \{1, 2, \ldots\}$*. Suppose that, for any* n*, our model for* $Y_1, \ldots, Y_n$ *is exchangeable:*

$$p(y_1, \ldots, y_n) = p(y_{\pi(1)}, \ldots, y_{\pi(n)})$$

for all permutations π *of* $\{1, \ldots, n\}$*. Then our model can be written as:*

$$p(y_1, \ldots, y_n) = \int \left\{ \prod_{i=1}^{n} p(y_i|\theta) \right\} p(\theta)\, d\theta,$$

for some parameters θ*, some prior distribution on* θ *and some sampling model* $p(y|\theta)$*. The prior and the sampling model depend on the form of the belief model* $p(y_1, \ldots, y_n)$*.*

These results can be summarised as follows:

$$\left.\begin{aligned} &Y_1,\dots,Y_n|\theta \text{ are i.i.d} \\ &\theta \sim p(\theta) \end{aligned}\right\} \iff Y_1,\dots,Y_n \text{ are exchangeable } \forall\, n.$$

1.5 Conjugacy

Here we explain the notion of *conjugacy*, and we refer to Hoff (2009), Chapter 3 for additional details.

The importance of these particular distributions is due to the fact that they provide a certain convenience in working with prior and posterior distributions. The posterior distribution, in fact, does not need to be established again, but will be like the prior one, with the parameters changed. Obtaining the posterior distribution reduces to updating the parameters of the prior distribution.

All this is done on the assumption that we are working with exchangeable observations.

Definition 1.6. A class $\mathscr{P}$ of prior distributions for θ is called *conjugate* for a sampling model $p(y_1,\dots,y_n|\theta)$ if

$$p(\theta) \in \mathscr{P} \Rightarrow p(\theta|y_1,\dots,y_n) \in \mathscr{P}.$$

Example 1.1. It can be demonstrated the conjugacy of the beta family for the Binomial sampling model. It results that if $\theta \sim \text{Beta}(a,b)$ and $Y|\theta \sim \text{Binomial}(m,\theta)$, then $\{\theta|Y=y\} \sim \text{Beta}(a+y, b+m-y)$.

Example 1.2. The conjugacy of the gamma family for the Poisson sampling model is also valid. If $\theta \sim \text{Gamma}(a,b)$ and $Y_1,\dots,Y_n|\theta \sim \text{Poisson}(\theta)$, then $\{\theta|Y_1,\dots,Y_n\} \sim \text{Gamma}(a+\sum_{i=1}^n Y_i, b+n)$.

1.5.1 Exponential families and conjugate priors

A one-parameter exponential family model is any model whose densities can be expressed as $p(y|\theta) = h(y)c(\theta)e^{\theta t(y)}$, where θ is the unknown parameter and $t(y)$ is the sufficient statistic that offers all the information needed to draw inferences about θ.

Diaconis & Ylvisaker (1979) investigate conjugate prior distributions for general exponential family models, specifically priors of the kind $p(\theta|n_0, t_0) = k(n_0, t_0)c(\theta)$. The induced posterior distribution is obtained by combining such prior information with information from $Y_1, \dots, Y_n \overset{\text{i.i.d}}{\sim} p(y|\theta)$:

$$\begin{aligned} p(\theta|y_1, \dots, y_n) &\propto p(\theta)p(y_1, \dots, y_n|\theta) \\ &\propto c(\theta)^{n_0+n}\exp\left\{\theta\left[n_0 t_0 + \sum_{i=1}^{n} t(y_i)\right]\right\} \\ &\propto p(\theta|n_0 + n, n_0 t_0 + n\bar{t}(\mathbf{y})), \end{aligned}$$

where $\bar{t}(\mathbf{y}) = \sum_{i=1}^{n} t(y_i)/n$.

The integer n_0 represents a "prior sample size", a measure of how informative the prior is, and t_0 represents a "prior estimation" of $t(Y)$.

1.6 The normal model

With references to Hoff (2009) Chapter 5, a random variable Y is normally distributed with mean μ and variance $\sigma^2 > 0$ if the density of Y is given by:

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}, -\infty < y < \infty \, .$$

The normal distribution is significant because of the central limit theorem, which states that the sum of a set of random variables is nearly normally distributed under extremely broad conditions. This suggests that the normal sampling approach will be acceptable for data resulting from the additive impacts of several factors.

1.6.1 Inference for the mean, conditional on the variance

It is supposed a model like $Y_1, \dots, Y_n | \mu, \sigma^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then the joint sampling density is given by

$$\begin{aligned} p(y_1, \dots, y_n | \mu, \sigma^2) = \prod_{i=1}^{n} p(y_i | \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\}. \end{aligned}$$

When we extend the quadratic term in the exponent, we observe that $p(y_1, \dots, y_n | \mu, \sigma^2)$ simply depends on $y_1, \dots, y_n$ through

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i^2 - 2\frac{\mu}{\sigma^2} \sum_{i=1}^{n} y_i + n\frac{\mu^2}{\sigma^2}.$$

This demonstrates that $\left\{ \sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} y_i \right\}$ constitute a two-dimensional sufficient statistic. From these quantities can be obtained the values of $\bar{y} = \sum_{i=1}^{n} y_i / n$ and $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})/(n-1)$. For this reason also $\{\bar{y}, s^2\}$ are a sufficient statistics. This two-parameter model's inference may be divided into two one-parameter issues. We will start with the challenge of inferring θ when σ^2 is known, and we will employ a conjugate prior distribution to do it. We observe that the posterior satisfies the following condition for any prior $p(\mu|\sigma^2)$:

$$\begin{aligned} p(\mu | y_1, \dots, y_n, \sigma^2) &\propto p(\mu|\sigma^2) e^{\frac{1}{2\sigma^2} \sum (y_i - \mu)^2} \\ &\propto p(\mu|\sigma^2) e^{c_1 (\mu - c_2)^2}. \end{aligned}$$

We can see that if $p(\mu|\sigma^2)$ is conjugate, it must contain quadratic terms such as $e^{c_1(\mu - c_2)^2}$. Thus, the normal family of probability densities on $\mathbb{R}$ is the simplest, implying that if $p(\mu|\sigma^2)$ is normal and $y_1, \dots, y_n$ are i.i.d normal(μ, σ^2), then $p(\mu | y_1, \dots, y_n, \sigma^2)$ is a normal density as well.

If $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$ then:

$$\begin{aligned} p(\mu|y_1,\dots,y_n,\sigma^2) &= \frac{p(\mu|\sigma^2)p(y_1,\dots,y_n|\mu,\sigma^2)}{p(y_1,\dots,y_n|\sigma^2)} \\ &\propto p(\mu|\sigma^2)p(y_1,\dots,y_n|\mu,\sigma^2) \qquad (1.4) \\ &\propto \exp\left\{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right\}\exp\left\{-\frac{n}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}. \end{aligned}$$

We can get from Equation (1.4) a function that has exactly the same shape as a normal density curve. Thus, we obtain:

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto \exp\left\{-\frac{1}{2}\left(\frac{\mu-\frac{b}{a}}{\frac{1}{\sqrt{a}}}\right)^2\right\},$$

where

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2},\ b = \frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}.$$

We refer to the mean and variance of this density as μ_n and τ_n^2 where

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \text{ and } \mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}.$$

The posterior parameters τ_n^2 and μ_n combine the prior parameters τ_0^2 and μ_0 with terms from the data. *Posterior variance and precision*:

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\ . \qquad (1.5)$$

The prior inverse variance is combined with the inverse of the data variance as it can be seen in the Equation (1.5) . Inverse variance is often referred to as the *precision*.

Let define some quantities:

- $\tilde{\sigma}^2 = 1/\sigma^2$ is the sampling precision;
- $\tilde{\tau}_0^2 = 1/\tau_0^2$ is the prior precision;
- $\tilde{\tau}_n^2 = 1/\tau_n^2$ is the posterior precision.

These imply that Equation (1.5) becomes:

$$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2.$$

Posterior mean. Notice that:

$$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}\ . \tag{1.6}$$

Note that $\tilde{\sigma}^2$ is the sampling precision. As a result of the Equation (1.6) the posterior mean is a weighted average of the prior mean and the sample mean. The weight on the sample mean is the sampling precision of the sample mean. The weight on the prior mean is $1/\tau_0^2$, the prior precision. If the prior mean were based on k_0 prior observations from the same population, we should put $\tau_0^2 = \sigma^2/k_0$, which is the variance of the prior mean. The formula for the posterior mean reduces to:

$$\mu_n = \frac{k_0}{k_0 + n}\mu_0 + \frac{n}{k_0 + n}\bar{y}\ .$$

1.6.2 Joint inference for the mean and the variance

Also with joint prior distributions $p(\mu, \sigma^2)$ for μ and σ^2, posterior inference proceeds using Bayes' rule:

$$p(\mu, \sigma^2 | y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n | \mu, \sigma^2) p(\mu, \sigma^2)}{p(y_1, \ldots, y_n)}\ .$$

A joint distribution for two quantities can be expressed as the product of a conditional probability and a marginal probability:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2) p(\sigma^2)\ .$$

We say that if σ^2 is known then a conjugate prior distribution for μ was $\mathcal{N}(\mu_0, \tau_0^2)$. In the particular case in which $\tau_0^2 = \sigma^2/k_0$:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2) = \text{dnorm}(\mu, \mu_0, \tau_0 = \frac{\sigma}{\sqrt{k_0}} p(\sigma^2),$$

the parameters μ_0 and k_0 can be interpreted as the mean and sample size from a set of prior observations. For the variance is needed a family of prior distributions that has support on $(0, \infty)$, such as the gamma family. However this family is not conjugate for the normal variance σ^2, but for the precision $1/\sigma^2$. For this reason, when we consider such a prior distribution, σ^2 has an *inverse-gamma* distribution:

$$precision = \frac{1}{\sigma^2} \sim \text{gamma}(a, b)$$
$$variance = \sigma^2 \sim \text{inverse-gamma}(a, b)$$

In particular, instead of using generic parameters as a and b, we will work with a prior distribution such as:

$$\frac{1}{\sigma^2} \sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right),$$

the parameters (σ_0^2, ν_0) can be interpret as the sample variance and sample size of prior observations.

Posterior inference

Suppose our prior distribution and sampling model are as follows:

$$\begin{aligned}
&\frac{1}{\sigma^2} \sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)\\
&\mu|\sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right)\\
&Y_1, \dots, Y_n|\mu, \sigma^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2).
\end{aligned}$$

Such as the prior distribution for μ and σ^2, also the posterior distribution can be decomposed:

$$p(\mu, \sigma^2|y_1, \dots, y_n) = p(\mu|\sigma^2, y_1, \dots, y_n)p(\sigma^2|y_1, \dots, y_n).$$

As a consequence of the previous considerations the conditional distribution of μ given the data and σ^2 is:

$$\begin{aligned}&\left\{\mu|y_1,\dots,y_n,\sigma^2\right\} \sim \mathcal{N}\left(\mu_n,\frac{\sigma^2}{k_n}\right)\\ &k_n = k_0 + n \ \text{ and } \ \mu_n = \frac{k_0\mu_0 + n\bar{y}}{k_n}.\end{aligned}$$

The posterior distribution of σ^2 can be obtained from:

$$\begin{aligned}p(\sigma^2|y_1,\dots,y_n) &\propto p(\sigma^2)p(y_1,\dots,y_n|\sigma^2)\\ &= p(\sigma^2)\int p(y_1,\dots,y_n|\mu,\sigma^2)p(\mu|\sigma^2)\,d\mu.\end{aligned}$$

Thus we get:

$$\begin{aligned}&\left\{\frac{1}{\sigma^2}|y_1,\dots,y_n\right\} \sim \text{gamma}\left(\frac{\nu_n}{2},\nu_n\frac{\sigma_n^2}{2}\right)\\ &\nu_n = \nu_0 + n\\ &\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_n}(\bar{y}+\mu_0)^2\right].\end{aligned}$$

The quantity ν_0 can be interpreted as a prior sample size, from which is obtained a prior sample variance of σ_0^2. Point out that $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$ is the sample variance.

Chapter 2

Algorithms for random sampling

After discussing the basic concepts of Bayesian statistics, this chapter will introduce the algorithms for random sampling. The Monte Carlo method will be discussed first. Indeed, this method is the simplest one and it is the basis of Markov Chain Monte Carlo (MCMC) algorithms, also discussed here.
The chapter will conclude with Gibbs sampling, the key object of this discussion, an example of which will be given later on.

The basic idea of these algorithms is to use randomness to solve problems that are, in principle, deterministic. In theory, Monte Carlo techniques can be used to solve any problem with a probabilistic interpretation. The integrals given by the expected value of a random variable can be approximated by taking the sample average of independent samples of the variable, according to the law of large numbers.

2.1 Monte Carlo approximation

We have already discussed the benefits of the conjugate priors for an unknown parameter θ. They are particularly helpful since a conjugate prior guarantees a posterior distribution for which there were simple formulas for posterior means and variances. Other aspects of a posterior distribution are frequently summarised. For example, for a random set A, we could be interested in calculating $\mathbb{P}(\theta \in A|y_1, \dots, y_n)$. We could also be interested in the means and standard deviations of some function of θ or the predictive distribution of missing or unobserved data.

The Monte Carlo approach may be used to estimate all of these posterior quantities of interest if we can generate random sample values of the parameters from their posterior distributions. For more details refer to Hoff (2009) Chapter 4.

2.1.1 The Monte Carlo method

Let θ be a parameter of interest and let $y_1, \dots, y_n$ be the numerical values of a sample from a distribution $p(y_1, \dots, y_n|\theta)$. Suppose we could sample S independent random θ-values from the posterior distribution $p(\theta|y_1, \dots, y_n)$:

$$\theta^{(1)}, \dots, \theta^{(S)} \overset{\text{i.i.d}}{\sim} p(\theta|y_1, \dots, y_n).$$

Then the empirical distribution of the samples $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ would approximate $p(\theta|y_1, \dots, y_n)$, where the approximation improves as S increases. The empirical distribution of $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ is known as a *Monte Carlo approximation* to $p(\theta|y_1, \dots, y_n)$.

The empirical distribution of the Monte Carlo samples provides an increasingly accurate approximation to the true density as S gets larger. Additionally, let $g(\theta)$ be any function. The law of large numbers says that if $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ are i.i.d. samples from $p(\theta|y_1, \dots, y_n)$, then:

$$\lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} g(\theta^{(s)}) = E[g(\theta)|y_1, \dots, y_n] = \int g(\theta)p(\theta|y_1, \dots, y_n)\, d\theta.$$

This implies that:

- $\lim_{S\to\infty} \bar{\theta} = \lim_{S\to\infty} \sum_{s=1}^{S} \theta^{(s)}/S = E[\theta|y_1, \dots, y_n]$

  The sample mean of the Monte Carlo samples, is approximately the true expected value;

- $\lim_{S\to\infty} \sum_{s=1}^{S} (\theta^{(s)} - \bar{\theta})^2/(S-1) = Var[\theta|y_1, \dots, y_n]$

  The Monte Carlo standard error is the approximation to the standard deviation;

- the empirical distribution of $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ is approximated with the *cumulative distribution function*.

The denominator S is chosen so that the Monte Carlo standard error is smaller than the precision selected to estimate $E[\theta|y_1,\dots,y_n]$.

2.1.2 Posterior inference for arbitrary functions

Suppose we are interested in the posterior distribution of some function $g(\theta)$ of θ.

The law of large numbers says that if we generate a sequence $\{\theta^{(1)},\dots,\theta^{(S)}\}$ from the posterior distribution of θ, then $\sum_{i=1}^{S} g(\theta^{(i)})/S$ converges to $E[g(\theta)|y_1,\dots,y_n]$. Other aspects of the posterior distribution of $g(\theta)$ can also be investigated.

It is necessary to use a Monte Carlo approach: for S times a θ value is independently sampled.

$$\begin{aligned}
&\text{sample } \theta^{(1)} \sim p(\theta|y_1,\dots,y_n), \text{ compute } g(\theta^{(1)})\\
&\text{sample } \theta^{(2)} \sim p(\theta|y_1,\dots,y_n), \text{ compute } g(\theta^{(2)})\\
&\qquad\qquad\vdots\\
&\text{sample } \theta^{(S)} \sim p(\theta|y_1,\dots,y_n), \text{ compute } g(\theta^{(S)}).
\end{aligned}$$

The sequence $\{g(\theta^{(1)}),\dots,g(\theta^{(S)})\}$ constitutes S independent samples from $p(g(\theta)|y_1,\dots,y_n)$, and so all the properties before described are valid.

2.1.3 Sampling from predictive distributions

An important feature of Bayesian inference is the existence of a *predictive distribution* for new observations.

Let $y_1,\dots,y_n$ be the outcomes from a sample of n random variables, and let $\tilde{Y}$ be an additional outcome from the same population that has yet to be observed. The *predictive distribution* of $\tilde{Y}$ is the conditional distribution of $\tilde{Y}$ given $\{Y_1 = y_1,\dots,Y_n = y_n\}$. For conditionally i.i.d variables this distribution can be derived from the distribution of $\tilde{Y}$ given θ and the posterior distribution of θ:

$$\begin{aligned}
\mathbb{P}(\tilde{Y} = \tilde{y}|y_1,\dots,y_n) &= \int \mathbb{P}(\tilde{Y} = \tilde{y}, \theta|y_1,\dots,y_n)\, d\theta\\
&= \int \mathbb{P}(\tilde{Y} = \tilde{y}|\theta, y_1,\dots,y_n) p(\theta|y_1,\dots,y_n)\, d\theta. \qquad (2.1)
\end{aligned}$$

The Equation (2.1) is called *posterior predictive distribution*, because it conditions on an observed dataset.

A predictive model that integrates over unknown parameters but is not conditional on observed data like $\mathbb{P}(\tilde{Y} = \tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)\,d\theta$ is called *prior predictive distribution*. Such a distribution can be useful for evaluating whether a prior distribution for θ actually translates into reasonable prior beliefs for the observable data $\tilde{Y}$.

We will be able to sample from $p(\theta|y_1, \dots, y_n)$ and $p(y|\theta)$ in many modelling situations, but $p(\tilde{y}|y_1, \dots, y_n)$ will be too intricate to sample directly. We may use a Monte Carlo technique to sample from the posterior predictive distribution indirectly.

Since $p(\tilde{y}|y_1, \dots, y_n) = \int p(\tilde{y}|\theta)p(\theta|y_1, \dots, y_n)\,d\theta$, we see that $p(\tilde{y}|y_1, \dots, y_n)$ is the posterior expectation of $p(\tilde{y}|\theta)$. To obtain the posterior predictive probability that $\tilde{Y}$ is equal to some specific value $\tilde{y}$, we sample $\theta^{(1)}, \dots, \theta^{(S)} \overset{\text{i.i.d}}{\sim} p(\theta|y_1, \dots, y_n)$ and then approximate $p(\tilde{y}|y_1, \dots, y_n)$ with $\sum_{s=1}^{S} p(\tilde{y}|\theta^{(s)})/S$. This procedure will work well if $p(y|\theta)$ is discrete and the quantities of interest are easily computed from this distribution. Obtaining a set of samples of $\tilde{Y}$ from its posterior predictive distribution can be done as follows:

$$\begin{aligned}
&\text{sample } \theta^{(1)} \sim p(\theta|y_1, \dots, y_n), \ \text{sample } \tilde{y}^{(1)} \sim p(\tilde{y}|\theta^{(1)}) \\
&\text{sample } \theta^{(2)} \sim p(\theta|y_1, \dots, y_n), \ \text{sample } \tilde{y}^{(2)} \sim p(\tilde{y}|\theta^{(2)}) \\
&\qquad\qquad \vdots \\
&\text{sample } \theta^{(S)} \sim p(\theta|y_1, \dots, y_n), \ \text{sample } \tilde{y}^{(S)} \sim p(\tilde{y}|\theta^{(S)}).
\end{aligned}$$

The sequence $\{(\theta, \tilde{y})^{(1)}, \dots, (\theta, \tilde{y})^{(S)}\}$ constitutes S independent samples from the joint posterior distribution of $(\theta, \tilde{Y})$, and the sequence $\{\tilde{y}^{(1)}, \dots, \tilde{y}^{(S)}\}$ constitutes S independent samples from the marginal posterior distribution of $\tilde{Y}$, which is the posterior predictive distribution.

It is important to note that the empirical distribution of sampled data does not always match the distribution of the population from which the data were generated, and in fact, if the sample size is small, it might appear quite different. Sample empirical distributions produced from a smooth population distribution can be rather irregular. It may be useful in such instances to have a predictive distribution that smooths out the bumps in the empirical distribution.

2.1.4 Markov Chain Monte Carlo methods

Previously, the parameters considered had the assumption of independence. We are now in the case where they are dependent and follow a Markov chain. A complete reference is the book of Robert & Casella (2004).

Definition 2.1 (Markov chain)**.** A sequence $Y^{(0)}, Y^{(1)}, \dots, Y^{(R)}$ of random elements is a *Markov chain* if:

$$\mathbb{P}(Y^{(r+1)} \in A | y^{(0)}, \dots, y^{(r)}) = \mathbb{P}(Y^{(r+1)} \in A | y^{(r)}).$$

When a Markov chain admits an invariant or stationary probability distribution, its level of stability increases. Invariant distribution means that the marginal distributions of $Y^{(r)}$ and $Y^{(r+1)}$ are the same and are equal to the probability density $p(y)$, although $Y^{(r)}$ and $Y^{(r+1)}$ remain dependent.

A stationary law is not included in every Markov chain. However, for sampling purposes, Markov chains should always converge to an invariant distribution. Indeed, the stationary distribution $p(y)$ in Markov Chain Monte Carlo reflects the goal density from which we want to simulate.

Then we will employ the following approximation:

$$\int g(y)p(y)\, dy \approx \frac{1}{R}\sum_{r=1}^{R} g(y^{(r)}),$$

where $y^{(1)}, \dots, y^{(R)}$ are generated according to a Markov chain, with $y^{(0)} \sim p(y)$. We assume that we will consider Markov chains that observes the following properties, that are informally defined as follows:

- *Irreducibility*

  The chain is irreducible if it does not lock in a local region of the sample space. In the discrete case the chain is irreducible if all states are connected.

- *Aperiodocity*

  The chain is aperiodic if it does not have any deterministic cycle.

- *Harris recurrent*

  In the discrete setting a state $j \in \mathbb{N}$ is recurrent if and only if the chain will eventually reach j with probability 1.

Although aperiodic and Harris recurrent Markov chains are highly stable, they do not necessarily admit an invariant distribution.

Definition 2.2 (Harris positive)**.** A Markov chain is said to be *Harris positive* if it is Harris recurrent and admits an invariant probability distribution.

Ergodic theorem

After introducing the essential arguments for the MCMC method, a fundamental theorem is presented here, which corresponds to the law of large numbers for Markov chains. It is the main justification for the use of MCMC methods. The following result holds independently on the initial conditions $Y^{(0)} \sim p_0$.

Theorem 2.1 (Ergodic Theorem)**.** *Let the Markov chain* $(Y^{(r)})_{r \geqslant 1}$ *be Harris positive with stationary distribution* p*. Let the function* g *be integrable with reference to* p*. Then:*

$$\lim_{R \to \infty} \frac{1}{R} \sum_{r=1}^{R} g(Y^{(r)}) = \int g(y)p(y)\, dy$$

almost surely.

2.1.5 Sampling the path of a Markov chain

Below there is a description of how to sample the Markov chain's path. Firstly it is simulated $Y^{(0)} \sim p_0$ and then the following values $(Y^{(r+1)}|Y^{(r)})$ according to the transition kernel. If a Markov chain has a stationary distribution p, simulating from it leads to a practical method for simulating from p as well. Moreover, the distribution p_r of $Y^{(r)}$ will eventually converge to the stationary law p we wish to simulate. Thus, $Y^{(B)}$ for $B > 0$ large enough can be regarded as a sample from p. The values $Y^{(1)}, Y^{(2)}, \ldots, Y^{(B)}$ represent the so-called *burn-in period*, that is the values the chain needs to reach convergence. These values should be discarded, but the choice of B is not so easy.

Hence, the approximations of functions of interest are based on the values:

$$\int g(y)p(y)\,dy \approx \frac{1}{R-B}\sum_{r=B+1}^{R} g(y^{(r)}),$$

which it relies on the Ergodic Theorem.

2.2 Gibbs sampling

There are several models for which sampling directly from the joint distribution is difficult. For such situations, it is more convenient to consider another distribution: the full-conditional of each parameter. In these cases, approximation of the posterior distribution can be done with the Gibbs sampler. This is an iterative algorithm that constructs a dependent sequence of values for parameters, whose distribution converges to the target joint posterior distribution.

This section will explain how the Gibbs sampler works in the context of normal models. See the Chapter 6 of Hoff (2009) for more details.

2.2.1 Semiconjugate prior distribution

In Section 1.6 we discussed the normal model and we characterised our uncertainty about μ as being dependent on σ^2:

$$\mu|\sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right).$$

In some situations we may want to specify our uncertainty about μ as being independent of σ^2, for that reason we will consider a distribution like:

$$\frac{1}{\sigma^2} \sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$$
$$\mu \sim \mathcal{N}\left(\mu_0, \tau_0^2\right).$$

These distributions are named *semiconjugate prior distribution*.

2.2.2 Full-conditional distributions

The aim of the algorithm is not to sample directly from the complete function, but from the conditional distribution of each parameter with respect to the remaining parameters and the observed data. These are the so-called *full conditional distributions*.

The posterior probability density can be written in the following way, which comes from the Bayes' rule:

$$p(\mu, \sigma^2|y_1, \dots, y_n) = p(\mu|\sigma^2, y_1, \dots, y_n)p(\sigma^2, y_1, \dots, y_n),$$

the same rule is valid also for the other parameters. For that reason sampling each full conditional distribution in turn, gives values that are proportional to the posterior distribution.

In our case, if $\{Y_1, \dots, Y_n|\mu, \sigma^2\} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$, it results that: $\{\mu|\sigma^2, y_1, \dots, y_n\} \sim \mathcal{N}(\mu_n, \tau_n^2)$ where:

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + n\frac{\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

Furthermore, for $1/\sigma^2$, which we named $\tilde{\sigma}^2$, the full conditional distribution given μ and $\{y_1, \dots, y_n\}$ is:

$$\begin{aligned} p(\tilde{\sigma}^2|\mu, y_1, \dots, y_n) &\propto p(y_1, \dots, y_n, \tilde{\sigma}^2, \mu) \\ &= p(y_1, \dots, y_n|\mu, \tilde{\sigma}^2)p(\mu|\tilde{\sigma}^2)p(\tilde{\sigma}^2). \end{aligned}$$

We supposed that μ and $\tilde{\sigma}^2$ are independent in the prior distribution, then $p(\mu|\tilde{\sigma}^2) = p(\mu)$ and

$$\begin{aligned} p(\tilde{\sigma}^2|\mu, y_1, \dots, y_n) \propto & \, p(y_1, \dots, y_n|\mu, \tilde{\sigma}^2)p(\tilde{\sigma}^2) \\ \propto & \left((\tilde{\sigma}^2)^{\frac{n}{2}} \exp\left\{-\tilde{\sigma}^2 \frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right\}\right) \cdot \\ & \left(\tilde{\sigma}^2)^{\frac{\nu_0}{2}-1} exp\left\{-\tilde{\sigma}^2 \nu_0 \frac{\sigma_0^2}{2}\right\}\right) \\ =&(\tilde{\sigma}^2)^{\frac{\nu_0}{2}-1} \exp\left\{-\tilde{\sigma}^2\left[\nu_0\sigma_0^2 + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right]\right\}. \end{aligned}$$

This is the form of a gamma density, and so

$$\{\sigma^2|\mu, y_1, \ldots, y_n\} \sim \text{Inverse-Gamma}\left(\nu_n/2, \nu_n\sigma_n^2(\mu)/2\right),$$

where

$$\nu_n = \nu_0 + n; \quad \sigma_n^2(\mu) = \frac{\nu_0\sigma_0^2 + ns_n^2(\mu)}{\nu_n}$$

and $s_n^2(\mu) = \sum_{i=1}^n (y_i - \mu)^2/n$, the unbiased estimate of σ^2 if μ were known.

2.2.3 Sampling from the full conditional distributions

As we have seen we can easily sample directly from $p(\sigma^2|\mu, y_1, \ldots, y_n)$, as well as from $p(\mu|\sigma^2, y_1, \ldots, y_n)$. However, we do not yet have a way to sample directly from the joint posterior distribution $p(\mu, \sigma^2|y_1, \ldots, y_n)$.

Gibbs sampler uses the relation between the full conditional distribution and the posterior distribution to suggest sampling each variable iteratively, while leaving the other variables in their previous state in time. When a full conditional is sampled in this way, a new value for the conditioned variable is picked and then immediately used to sample the other variables that have not been sampled yet. Suppose we were given $\sigma^{2(1)}$, a single sample from the marginal posterior distribution $p(\sigma^2|y_1, \ldots, y_n)$. Then we could sample:

$$\mu^{(1)} \sim p(\mu|\sigma^{2(1)}, y_1, \ldots, y_n)$$

and $\{\mu^{(1)}, \sigma^{2(1)}\}$ would be a sample from the joint distribution of $\{\mu, \sigma^2\}$. Additionally, $\{\mu^{(1)}, \sigma^{2(1)}\}$ can be considered a sample from the marginal distribution $p(\mu|y_1, \ldots, y_n)$. From this μ-value, we can generate

$$\sigma^{2(2)} \sim p(\sigma^2|\mu^{(1)}, y_1, \ldots, y_n).$$

But since $\mu^{(1)}$ is a sample from the marginal distribution of μ, and $\sigma^{2(2)}$ is a sample from the conditional distribution of σ^2 given $\mu^{(1)}$, then $\{\mu^{(1)}, \sigma^{2(2)}\}$ is also a sample from the joint distribution of $\{\mu, \sigma^2\}$. This generates samples iteratively with a Markov approach, in which the samples tend towards the posterior density ones.

More precisely given a current state of the parameters $\varphi^{(s)} = \{\mu^{(s)}, \sigma^{2(s)}\}$, we generate a new state as follows:

- sample $\mu^{(s+1)} \sim p(\mu|\tilde{\sigma}^{2(s)}, y_1, \dots, y_n)$;
- sample $\tilde{\sigma}^{2(s+1)} \sim p(\sigma^2|\mu^{(s+1)}, y_1, \dots, y_n)$;
- let $\varphi^{(s+1)} = \{\mu^{(s+1)}, \tilde{\sigma}^{2(s+1)}\}$.

This algorithm is called Gibbs sampler, and generates a dependent sequence of our parameters $\{\varphi^{(1)}, \varphi^{(2)}, \dots, \varphi^{(S)}\}$.

It is important to note that it is a Markov chain: it is a chain of dependent values, in which $\varphi^{(s)}$ depends on $\varphi^{(0)}, \dots, \varphi^{(s-1)}$ only through $\varphi^{(s-1)}$, i.e. $\varphi^{(s)}$ is conditionally independent of $\varphi^{(0)}, \dots, \varphi^{(s-2)}$ given $\varphi^{(s-1)}$.

Since it is a specific MCMC method, the following properties are valid:

$$\lim_{s\to\infty} \mathbb{P}(\varphi^{(s)} \in A) \to \int_A p(\varphi)\, d\varphi$$

with A a generic measurable set. This means that the *sampling distribution* of $\varphi^{(s)}$ approaches the *target distribution* as $s \to \infty$, no matter what the starting value $\varphi^{(0)}$ is. Furthermore, for most functions g of interest:

$$\lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} g(\varphi^{(s)}) \to E[g(\varphi)] = \int g(\varphi)p(\varphi)\, d\varphi.$$

We can approximate $E[g(\varphi)]$ with the sample average of $\{g(\varphi^{(1)}), \dots, g(\varphi^{(S)})\}$. To be a good approximation for a wide range of functions g, we need the empirical distribution of the simulated sequence $\{\varphi^{(1)}, \dots, \varphi^{(S)}\}$ to look like the target distribution $p(\varphi)$.

As we have already explained, the Markov chain produced by the Gibbs sampler possibly begins to converge after the generation of many samples. Thus, it is good practice to discard the first k produced values, where k depends on the speed of convergence of the chain. These are usually referred to as *burn–in* iterations.

An example of the Gibbs Sampler algorithm will be given in Chapter 4 of this discussion.

Chapter 3

Finite Mixture Models

Mixture models can be employed in situations when the population of sampling units is divided into several sub-populations, each having its own simple model. This chapter will introduce these models and more insights can be found in Chapter 22 of Gelman et al. (2015) and in Green (2018).

The essential idea behind mixture models is to incorporate unseen random variables, commonly labelled as a vector or matrix $\boldsymbol{z}$, that indicate the mixture component from which each specific observation is chosen. As a result, a mixture model has a hierarchical representation; the observed variables y are conditionally modelled on the vector $\boldsymbol{z}$, and the vector $\boldsymbol{z}$ is given a probabilistic specification. It is sometimes helpful to consider the mixing indicators as missing data. Doing the average across the distribution of the indicator variables yields inferences about quantities of interest, such as parameters inside the probability model for y. This means pulling $(\theta, \boldsymbol{z})$ from their joint posterior distribution in the simulation framework.

3.1 The set up and the interpretation of mixture models

Suppose we want to model the distribution of a random sample $\boldsymbol{y} = (y_1, \dots, y_n)$ as a mixture of H components. It is considered that it is unknown which component of the mixture underlies any specific observation. The h-th component distribution, $f_h(y_i|\boldsymbol{\theta}_h)$, is considered to depend on a parameter vector $\boldsymbol{\theta}_h$ for $h = 1, \dots, H$ and on the parameter λ_h expressing the proportion of the population from component h, which is a non-negative value with $\sum_{h=1}^{H} \lambda_h = 1$. It is common to suppose that the mixing components all belong to the same parametric family, such as normal, but have distinct parameter vectors. In that situation, the sampling distribution of $\boldsymbol{y}$ is:

$$p(y_i|\boldsymbol{\theta},\boldsymbol{\lambda}) = \lambda_1 f(y_i|\boldsymbol{\theta}_1) + \lambda_2 f(y_i|\boldsymbol{\theta}_2) + \cdots + \lambda_H f(y_H|\boldsymbol{\theta}_H) = \sum_{h=1}^{H} \lambda_h f(y_i|\boldsymbol{\theta}_h).$$

In other words in mixture models, the component density f is fixed and known, the component-specific parameters $\boldsymbol{\theta}_h$ and the weights λ_h are usually considered to be unknown and H is also sometimes unknown. It can be useful to allow slightly more generality:

$$p(y_i|\boldsymbol{\theta},\boldsymbol{\lambda}) = \sum_{h=1}^{H} \lambda_h f_h(y_i|\boldsymbol{\theta}_h),$$

where different components are allowed to to have different parametric forms. Modern inference for mixture models almost always uses the likelihood function, which in the case of n independently and identically distributed observations from a mixture model has the form:

$$p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\lambda}) = \prod_{i=1}^{n} \sum_{h=1}^{H} \lambda_h f(y_i|\boldsymbol{\theta}_h). \tag{3.1}$$

Many problems with mixture models derive from this product-of-sums.

3.1.1 Latent allocation variables

Suppose the population from which we are sampling is heterogeneous: there are multiple groups, indexed by $h = 1, 2, \ldots, H$, present in the population in proportions λ_h. When sampling from group h, observations are assumed drawn from density $f(\cdot|\boldsymbol{\theta}_h)$. Thus, in every mixture model there is an unobserved indicator variable z_{ih}, with

$$z_{ih} = \begin{cases} 1 & \text{if the } i\text{-th unit is drawn from the } h\text{-th mixture component} \\ 0 & \text{otherwise.} \end{cases}$$

Then we can imagine that an observation y drawn from the population is realised in two steps: first, the group z is drawn from the index set $h = 1, 2, \ldots, H$, with $\mathbb{P}(z_h = 1) = \lambda_h$; and secondly, given z, y is drawn from $f(\cdot|\boldsymbol{\theta}_h)$.

The $\boldsymbol{z}_i$ are *latent* random variables, they are usually called *allocation variables* in the mixture model context.

Given $\boldsymbol{\lambda}$, Multin$(1; \lambda_1, \ldots, \lambda_H)$ is the distribution of each vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})$. In this case, the mixing parameters are regarded as hyperparameters that determine the distribution of $\boldsymbol{z}$. The conditional joint distribution of the observable data $\boldsymbol{y}$ and the unseen indicators $\boldsymbol{z}$ can be stated:

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\lambda}) = p(\boldsymbol{z}|\boldsymbol{\lambda})p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{h=1}^{H} (\lambda_h f(y_i|\boldsymbol{\theta}_h))^{z_{ih}}, \tag{3.2}$$

with exactly one of z_{ih} equaling 1 for each i. This is the *complete data likelihood*. The number of the mixture components H, is assumed to be known and fixed. The finite mixture is a special case of the more general specification $p(y_i) = \int p(y_i|\boldsymbol{\theta})\boldsymbol{\lambda}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

3.1.2 Some possible difficulties with mixture models

The model parameters are not recognised if more than one option of the likelihood function is obtained. All finite mixture models are non-identifiable in one sense; if the group labels are permuted, the distribution remains unchanged. For many problems, an informative prior distribution has the effect of identifying specific components with specific sub-populations. The prior distribution for the finite mixture model parameters $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is taken, in most applications, to be a product of independent prior distributions on $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. The natural conjugate prior distribution is the Dirichlet, $\boldsymbol{\lambda} \sim$ Dirichlet$(\alpha_1, \cdots, \alpha_H)$, if the vector of mixture indicators $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})$ is treated as a multinomial with parameter $\boldsymbol{\lambda}$. The mean of the prior distribution for $\boldsymbol{\lambda}$ is described by the relative sizes of the Dirichlet parameters α_h, and the sum of the α_h's is a measure of the prior distribution's strength. The generic Dirichlet distribution is:

$$f(x_1, \ldots, x_H|\alpha_1, \ldots, \alpha_H) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_H)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_H)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_H^{\alpha_H - 1}$$

We use $\boldsymbol{\theta}$ to represent the vector consisting of all of the parameters in the mixture components, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H)$. For now we do not make any assumptions about the prior distribution $p(\boldsymbol{\theta})$.

An improper non-informative prior distribution for $\boldsymbol{\lambda}$, corresponding to $\alpha_i = 0$, may cause a problem if the data do not indicate that all H components are present in the sample. When inappropriate prior distributions are used for the component parameters, issues are more likely to occur.

3.1.3 Posterior modes using EM

For finite mixture models there is often uncertainty concerning the number of mixture components H to include in the model. Although computing models with high values of H might be expensive, it is preferable to start with a small mixture and test the fit. The posterior predictive distribution of an appropriate test quantity can be used to see if the present number of components adequately describes the range of observed data. Each of the $i = 1, \dots, n$ items in the sample belongs to one of H sub-populations, with each latent sub-population or latent class having a distinct value for one or more parameters in a parametric model. Let $z_i \in \{1, \dots, H\}$ denote the sub-population index for item i, with this index commonly referred to as the latent class status. Then, the response y_i for item i conditionally on z_i has the distribution

$$y_i | z_i \sim f(\cdot | \boldsymbol{\theta}_{z_i}).$$

In marginalising out the latent class status, assuming that the fraction of the population belonging to sub-population h equals $\mathbb{P}(z_i = h) = \lambda_h$, the following probability is obtained:

$$p(\boldsymbol{y} | \boldsymbol{\lambda}, \boldsymbol{\theta}) = \sum_{h=1}^{H} \lambda_h f(\boldsymbol{y} | \boldsymbol{\theta}_h),$$

which corresponds to a finite mixture with H components, with component h assigned probability weight λ_h. The most common form of inference is maximum likelihood, which is based on maximising of Equation (3.1), again using numerical methods. Such an approach is immediately interesting for lots of reasons, including its relevance when the basic model is developed, such as through the inclusion of covariates, assuming the numerical issues are resolved. The usual numerical approach to maximum likelihood estimation of mixture models uses the *EM algorithm*.

The EM algorithm can be used to estimate the parameters of a finite mixture model, averaging over the indicator variables. If the latent allocation variable z_i were known, we would have separate independent samples for each component h so that it would be a simple practice to estimate p. If the parameters $\boldsymbol{\theta}_h$ were known, instead, the allocation of the observations y_i to the different components could be done by choosing $z_i = h$ to maximise $f(y_i|\boldsymbol{\theta}_h)$. The E-step frequently requires the computation of the expected value of the sufficient statistics of the joint model of $(\boldsymbol{y}, \boldsymbol{z})$. This is done using the log of the complete-data likelihood, defined in the Equation (3.2), conditional on the last guess of the value of the mixture component parameters $\boldsymbol{\theta}$ and the mixture proportions $\boldsymbol{\lambda}$. In finite mixtures this is equivalent to compute the conditional expectation of the indicator variables by Bayes' rule. We suggest choosing a fairly large number starting points by simplifying the model or random sampling. Each z_{ih} is replaced by its conditional expectation $E(z_{ih}|\boldsymbol{\theta}, \boldsymbol{\lambda}, y_i) = \tau_{ih}$ given the current values of the parameters and weights that is:

$$\tau_{ih} = \frac{\lambda_h f(y_i|\boldsymbol{\theta}_h)}{\sum_{h'=1}^{H} \lambda_{h'} f(y_i|\boldsymbol{\theta}_{h'})}.$$

In the M-step, λ_h and $\boldsymbol{\theta}_h$ are updated to maximise the corresponding expected complete-data log likelihood,

$$\sum_{i=1}^{n} \sum_{h=1}^{H} \tau_{ih} \log(\lambda_h f(y_i|\boldsymbol{\theta}_h)).$$

When the λ_h and $\boldsymbol{\theta}_h$ vary independently for each component h, this maximisation may be done separately for each one.

3.1.4 Posterior simulation using the Gibbs sampler

Starting values for the Gibbs sampler can be derived from an appropriate approximation to the posterior via relevance re-sampling. Given the current values of $\{\lambda_h\}$ and $\{\boldsymbol{\theta}_h\}$ the Gibbs sampler update for the allocation variables z_{ih} is to draw them independently from $\mathbb{P}(z_{ih} = 1|\boldsymbol{\theta}, \boldsymbol{\lambda}, y_i) = \tau_{ih}$, where τ_{ih} is defined in the E-step Section 3.1.3; this is simply Bayes' rule. The algorithm alternates two key steps for mixture models: drawing from the distribution of the indicators given the model parameters and drawing from the model parameters given the indicators. To update all of the model parameters, the second step may include numerous stages. Using conjugate families as prior distributions might be useful. In finite mixture models, obtaining draws from the distribution of the indicators, is typically simple: these are multinomial draws. Modeling problems, such as incorrectly applying a prior density, are frequently discovered during the iterative simulation stage of calculations. For example, a Gibbs sequence started near zero variance may never leave the area.

By ignoring the drawn indications once the Gibbs sampler has reached approximate convergence, posterior inferences about model parameters can be made. Each observation is chosen from the posterior distribution of the indicator variables, which provides information about the probable components.

3.1.5 Label switching and posterior computation

If there is interest in inferences on mixture component-specific parameters and clustering due to identifiability issues such as the so-called label ambiguity and label switching problem, it makes a significant difference in conducting the analysis and defining priors. The label ambiguity problem refers to the issue for which there is nothing in the likelihood to distinguish mixture component h as different from h'. When employing the EM method to maximise the probability of a finite mixture model, this problem becomes evident. If the method converges to a point estimate $(\hat{\lambda}_1, \dots, \hat{\lambda}_H)$, $(\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_H)$, there will be another estimate with an identical likelihood $(\hat{\lambda}_{K_1}, \dots, \hat{\lambda}_{K_H})$, $(\hat{\boldsymbol{\theta}}_{K_1}, \dots, \hat{\boldsymbol{\theta}}_{K_H})$, where $(K_1, \dots, K_H)$ is any permutation of the indexes $1, \dots, H$. A joint prior distribution for $(\lambda_1, \dots, \lambda_H)$, $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H)$ is required in a Bayesian model. The marginal posterior distribution of $\boldsymbol{\theta}_h$ will be similar for all h if the mixture components $1, \dots, H$ are exchangeable in the prior distribution. As a result, it is impossible to estimate a posterior distribution for mixture component h without distinguishing it from the rest.

A typical exchangeable prior would let

$$(\lambda_1, \dots, \lambda_H) \sim \text{Dirichlet}(a, \dots, a) \ \text{ and } \ \boldsymbol{\theta}_h \overset{\text{i.i.d}}{\sim} P_0,$$

independently, with P_0 an arbitrary common prior from which the mixture component-specific parameters are drawn.

We said that due to excheangeability of the mixture components, the marginal posterior distribution of $\boldsymbol{\theta}_h$ is identical for all $h \in \{1, \dots, H\}$ and hence the chains for each of the mixture component parameters have the same target distribution. For example, supposed a mixture of two Gaussians with means μ_1 and μ_2 in which one mixture component is located at $\mu = 0$ and the other component is located at $\mu = -1$. The Gibbs sample for μ_1 should then randomly jump between values close to 0 and values close to -1 if mixing is good. The posterior for μ_h has a multimodal form with one mode close to 0 and one close to -1. Due to this mode are well separated and there is a region of low probability density between the modes, then the Gibbs sampler will remain stuck for long intervals in one mode. For example, for the first 5000 iterations the μ_1 samples may be close to 0 and μ_2 samples close to -1.

It is important to avoid choosing a P_0 that is improper as the results may be sensitive to the size of the variance chosen. Instead, best results are obtained when P_0 is chosen to generate mixture components that are close to the support of the data. One way to accomplish this in practice is normalising the data in advance of the analysis to facilitate selection on P_0. An alternative way is to select the hyperparametrs in P_0 based on one's prior knowledge about the location and scale of the data.

3.1.6 Clustering and classification

Mixture models can be adopted by thinking of the data as clustered. The term cluster commonly suggests a degree of homogeneity within a cluster and a degree of separation between clusters, but it does not imply any specific within-cluster distribution.

Since the supposed parametric form of the component densities is incorrect, a fitted mixture model may require more components than there are apparent clusters in the data. Each homogeneous cluster may require many components to fit it well. Nevertheless, finite mixture modelling does provide one of the few simple and rigorous model-based approaches to clustering. It appears that a second level of indexing of weights and parameters is required for reasonable model-based inference about clusters using a within-cluster data distribution that is likely to be a mixture as

$$y \sim \sum_{h=1}^{H} \sum_{g=1}^{G_h} \lambda_{hg} f(\cdot|\boldsymbol{\theta}_{hg}),$$

where h indexes clusters and g components within clusters. The request of homogeneity within clusters and separation between clusters can be satisfied by appropriately modelling the $\boldsymbol{\theta}_{hg}$.

Chapter 4

Applications of the Gibbs sampler to mixture models

This chapter will show how the Gibbs sampler algorithm may be used with Gaussian mixture models. In particular, the first section will focus on simulated theoretical example in which the function of the algorithm implemented in R software will be defined and shown. Then, some examples will be provided using two datasets that are currently available in R: *Old Faithful Geyser Data* a base dataset of R and *Galaxy Data* from the *bmixture* library. Finally with *Concrete Compressive Data*, a real dataset, will be tested.

4.1 Construction of the algorithm

The starting example was taken from Chapter 7 of Robert & Casella (2010). Consider a normal mixture of two components that has the same variance and fixed weights

$$p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2).$$

We assume in addition a normal prior distribution $\mathcal{N}(0, \nu^2\sigma^2)$, with ν^2 known, on both means μ_1 and μ_2. The latent variables Z_i are defined as

$$\mathbb{P}(Z_i = 1) = 1 - \mathbb{P}(Z_i = 2) = p \text{ and } X_i|Z_i = k \sim \mathcal{N}(\mu_k, \sigma^2).$$

In other words, $Z_i \sim \mathrm{Bern}(p)$. Then, the complete data likelihood is:

$$p(\mu_1, \mu_2, \boldsymbol{z}|\boldsymbol{y}) \propto \exp\left\{-\frac{\mu_1^2 + \mu_2^2}{\nu^2\sigma^2}\right\} \cdot$$
$$\prod_{i:\, z_i=1} p\exp\left\{-\frac{(y_i - \mu_1)^2}{2\sigma^2}\right\} \prod_{i:\, z_i=2} (1-p)\exp\left\{-\frac{(y_i - \mu_2)^2}{2\sigma^2}\right\},$$

from which we can easily see that the full conditional distributions for the means are:

$$p(\mu_j|\boldsymbol{y}, \boldsymbol{z}, \sigma^2) \propto \exp\left\{-\frac{\mu_j^2}{2\nu^2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i:\, z_i=j} (y_i - \mu_j)^2\right\}$$

with $j = 1, 2$ and $p_1 = p$, $p_2 = 1 - p$. As a consequence we get:

$$\mu_j|\boldsymbol{y}, \boldsymbol{z}, \sigma^2 \sim \mathcal{N}\left(\frac{\nu^2}{1 + n_j\nu^2} \sum_{i:\, z_i=j} y_i, \frac{\sigma^2\nu^2}{1 + n_j\nu^2}\right) \tag{4.1}$$

with n_j the number of z_i that are equal to j.

For the latent variables the full conditional distribution is:

$$\mathbb{P}(Z_i = j|y_i, \mu_1, \mu_2) = \frac{p_j\exp\left\{-\frac{(y_i-\mu_j)^2}{2\sigma^2}\right\}}{\sum_{j'\in\{1,2\}} p_{j'}\exp\left\{-\frac{(y_i-\mu_{j'})^2}{2\sigma^2}\right\}}. \tag{4.2}$$

Then, we consider a simulated dataset $\boldsymbol{y}$ of 500 points from the $0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(2.7, 1)$ distribution. We want to improve the Gibbs sampler under this condition. The algorithm we have implemented is shown below:

Algorithm 1: Gibbs sampler for means

```
Load: the dataset, $\sigma^2$, $\nu^2$, p
Initialize: $\boldsymbol{z}$
for $i = 1, \dots, N_{iter}$ do
    sample $\mu_1$ from Equation (4.1);
    sample $\mu_2$ from Equation  (4.1);
    sample $\boldsymbol{z}$ from Equation (4.2)
Result: A vector containing the $N_{iter}$ values of $\mu_1$ sampled and another
        containing the $N_{iter}$ values of $\mu_2$ sampled
```

With this algorithm 1500 values for the means were sampled, where we have selected the following values: $p = 0.7$, $\sigma^2 = 0.01$ and $\nu^2 = 10$. Figure 4.1 represents the plot of the result over the log-posterior surface:

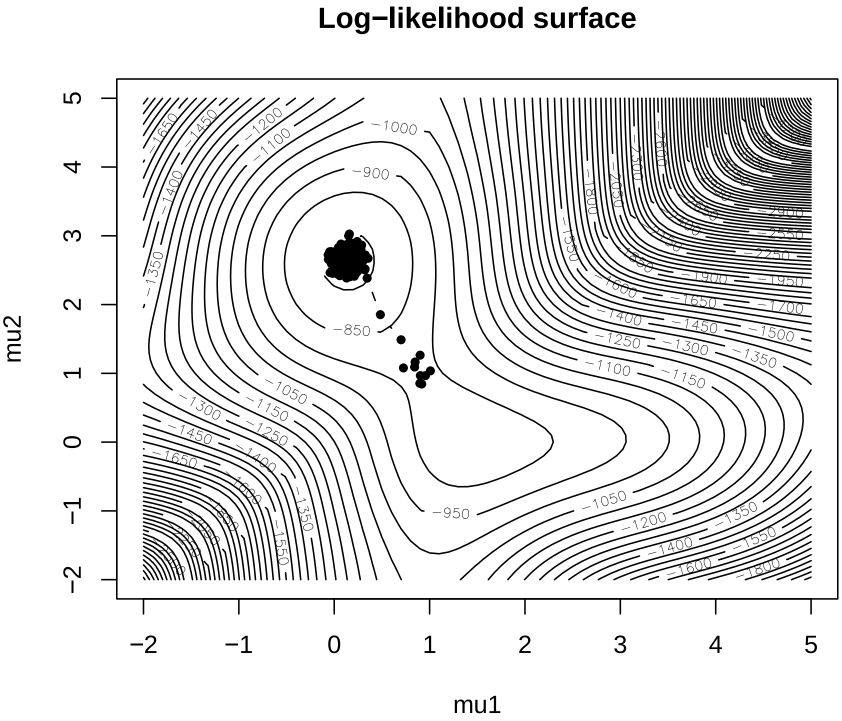


Figure 4.1: Gibbs sample of 1500 points for the mixture posterior against the log-posterior surface.

In order to make the algorithm more general, assume that σ^2 is no longer known and fixed, and also the number of components is an arbitrary value, $k = 1, 2, \ldots, n$. Consider a value σ_j^2 for each component of the mixture, where $j = 1, 2, \ldots, k$. Assume respectively such a prior distributions InvGamma(α_j, β_j). The distribution for the latent elements is specified as follows:

$$Z_i|\boldsymbol{p} \sim \text{Multinom}_k(1, \boldsymbol{p})$$
$$\boldsymbol{p} \sim \text{Dir}(\boldsymbol{\omega}).$$

Then the full conditionals for the means are:

$$\mu_j|\boldsymbol{y}, \boldsymbol{z} \sim \mathcal{N}\left(\frac{\nu^2}{\sigma_j^2 + n_j\nu^2}\sum_{i:\, z_i = j} y_i, \frac{\sigma_j^2\nu^2}{\sigma_j^2 + n_j\nu^2}\right). \tag{4.3}$$

As for the variances, we have:

$$\sigma_j^2|\boldsymbol{y},\boldsymbol{z} \sim \text{InvGamma}\left(\alpha_j - \frac{n_j}{2}, \beta_j + \frac{1}{2}\sum_{i:\, z_i=j}(y_i-\mu_j)^2\right), \tag{4.4}$$

assuming that the parameter α_j is equal for all the k components, and also for the parameter β_j this assumption is satisfied. For the weights:

$$\boldsymbol{p} \sim \text{Dir}(\boldsymbol{\omega}+\boldsymbol{n}) \tag{4.5}$$

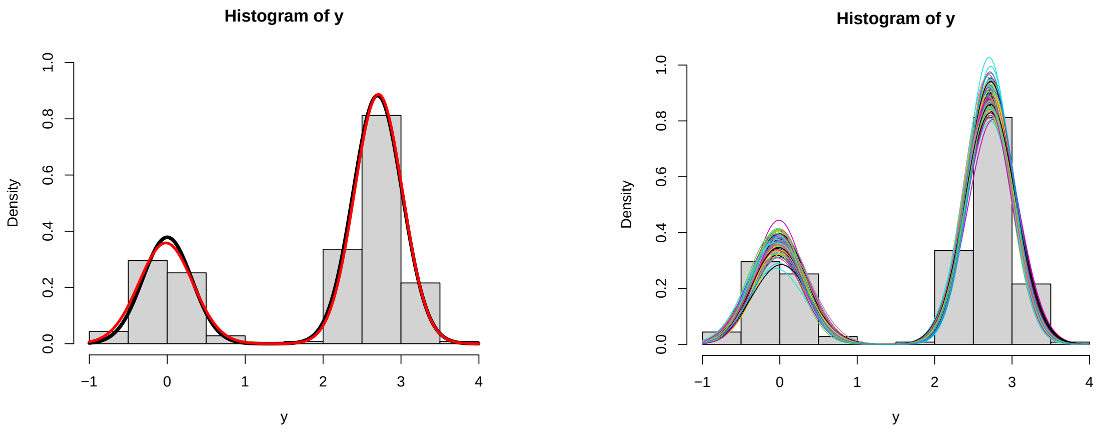
where $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_k)$ and $\boldsymbol{n} = (n_1, n_2, \dots, n_k)$, which is equivalent to $p_1 \sim \text{Beta}(\omega_1 + n_1, \omega_2 + n_2)$ and $p_2 = 1 - p_2$ when we consider two components. The Dirichlet distribution is a generalisation of the Beta distribution and describes the posterior parameters of a multinomial distribution of an observation. The Dirichlet distribution with two parameters is exactly the Beta distribution.

The Gibbs sampler is changed as follows:

```
Algorithm 2: Gibbs sampler for means and variances
  Load: the dataset, the support, $\alpha_1$, $\beta_1$, $\nu^2$, $k$
  Initialize: $\boldsymbol{z}$, $\boldsymbol{\sigma^2}$, $burn_{in}$
  for $i = 1, \dots, (burn_{in} + N_{iter})$ do
      for $j = 1, \dots, k$ do
          sample $\mu_j$ from Equation (4.3);
          sample $\sigma_j^2$ from Equation (4.4).
      sample $\boldsymbol{p}$ from Equation (4.5).
  Result: A matrix $(N_{iter} \times j)$ containing the $N_{iter}$ values of $\mu_j$ sampled, one
          of dimension $(N_{iter} \times j)$ containing the $N_{iter}$ values of $\sigma_j^2$ sampled,
          another of dimension $(N_{iter} \times j)$ containing the $N_{iter}$ values of $p$
          sampled and a matrix $(N_{iter} \times \#support)$ containing the values
          density estimated for the support with the parameters considered
          for all the $N_{iter}$ iterations
```

Since we are considering a theoretical example, it is possible to evaluate the goodness of the estimated density with the true one.
For each iteration, a density estimate can be obtained with the estimated parameters. The average of all estimated densities after the burn-in period is compared with the true density.

Let consider a mixture model with two components and the following distribution: $0.3\mathcal{N}(0, 0.1) + 0.7\mathcal{N}(2.7, 0.1)$ from which the dataset of 500 observations is sampled. The support of this distribution is $[-1, 4]$. The parameter ν^2 is initialized as 10. It is considered a burn-in period of 2300 iterations and then another 1100 iterations. The result obtained applying the algorithm is as follows:

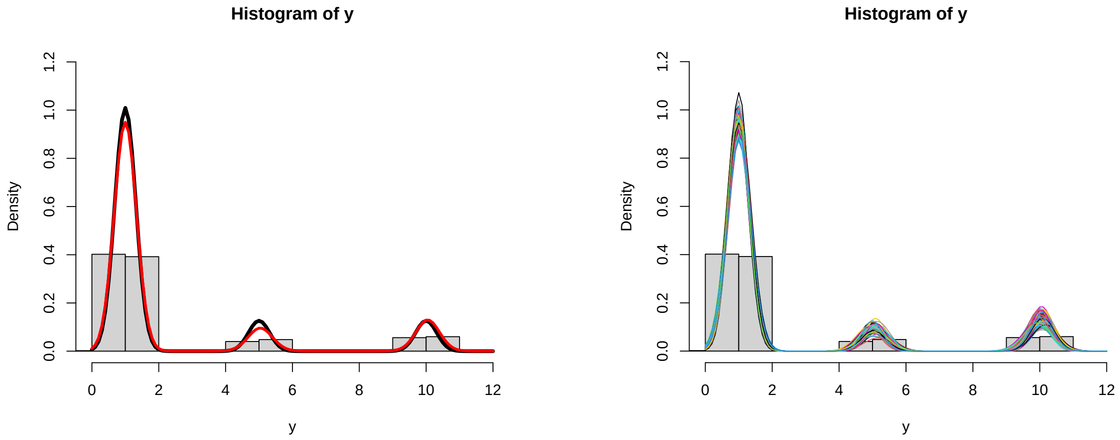


(a) Density estimated in red compared with the true density in black

(b) Densities estimated in the last 110 iterations

Figure 4.2: Density estimation for 2 components

Suppose a model with three components and a distribution like: $0.8\mathcal{N}(1, 0.1) + 0.1\mathcal{N}(5, 0.1) + 0.1\mathcal{N}(10, 0.1)$. It is sampled a dataset of 500 observations. The support is $[0, 12]$. Once again $\nu^2 = 10$ and 3400 iterations are performed of which 2300 constitute burn-in period. The result obtained is as above:



(a) Density estimated in red compared with the true density in black

(b) Densities estimated in the last 110 iterations

Figure 4.3: Density estimation for 3 components

After implementing the code on a theoretical example and verifying its operation, now we will apply the algorithm to two datasets provided by R to estimate their density.

4.2 Application of the Gibbs sampler to R datasets

The initial examples of datasets have been taken from R because they are suitable for use as tests as they are known and treated to be easily employed.

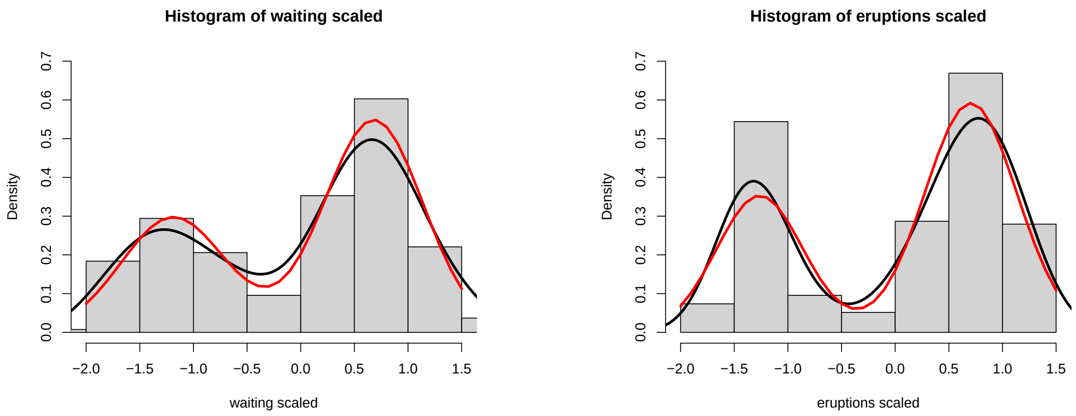
4.2.1 Old Faithful Geyser Data

As we read in RDocumentation (2022b), the Old Faithful Geyser Data is a data frame with 272 observations on 2 variables. The first variable represents the waiting time between eruptions, the second the duration of the eruption, considering the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Both variables are numeric.

Looking at the histogram of both variables, *waiting* and *eruptions*, a mixture model with two components is assumed for both. For further confirmation, a model-based clustering specific to finite mixture models of normals was used, provided by the *Mclust* library. The output obtained confirms the presence of two clusters.

The Gibbs sampler is used to estimate the respective density of the variables. It is considered the following values for the parameters for *eruptions*: $k = 2$, $\alpha_j = 1$ and $\beta_j = 5$, $\omega_j = 0.5$ for all $j = 1, \dots, k$ and $\nu^2 = 10$. Otherwise, for *waiting* it is used $\alpha_j = 15$.

These are the results obtained:

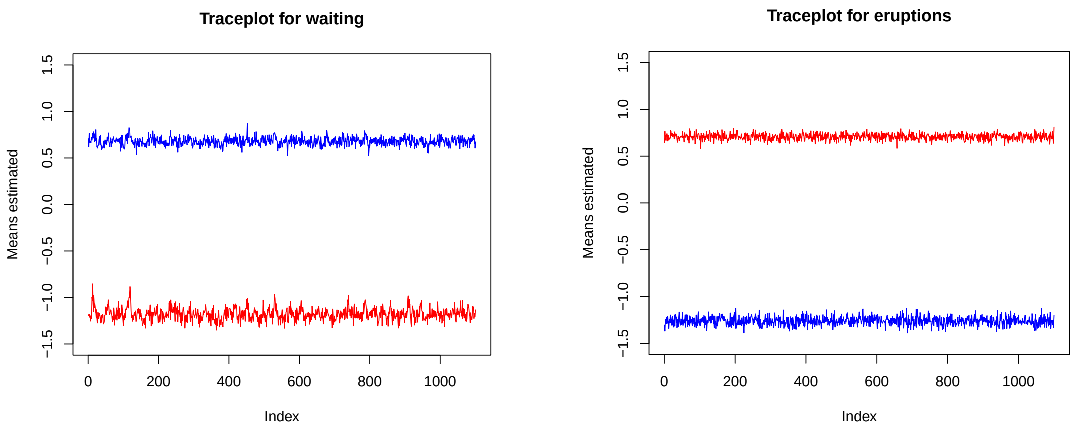


(a) Density estimated for waiting scaled in red with the Gibbs sampler, in black with the classic inference

(b) Density estimated for eruptions scaled in red with the Gibbs sampler, in black with the classic inference

Figure 4.4: Density estimation for the variables of Old Faithful Geyser Data

Looking at the traceplot of the averages confirms the convergence of the algorithm and the relevance of the averages identified. Distinct and appropriate values were identified.



(a) Traceplot of the avarages of waiting

(b) Traceplot of the avarages of eruptions

Figure 4.5: Traceplot of means

4.2.2 Galaxy Data

The library *bmixture* contains the Galaxy dataset, more information on which can be found at this link RDocumentation (2022a).

This dataset considers 82 observations of the velocities (in 1000 km/second) of distant galaxies diverging from our own, from six well-separated conic sections of the Corona Borealis. It contains only one variable named *speed*.

The same procedure adopted with the Old Faithful Geyser Data, will be used here. Four components are assumed through graphic histogram analysis, and the same number is obtained with Mclust. Although the number of components is bigger, the Gibbs sampler still works well as we can see in Figure 4.6. The parameters assume these values: $\nu^2 = 10$, $\alpha_j = 50$ and $\beta_j = 70$, $\omega_j = 0.5$ for all $j = 1, \ldots, 4$.

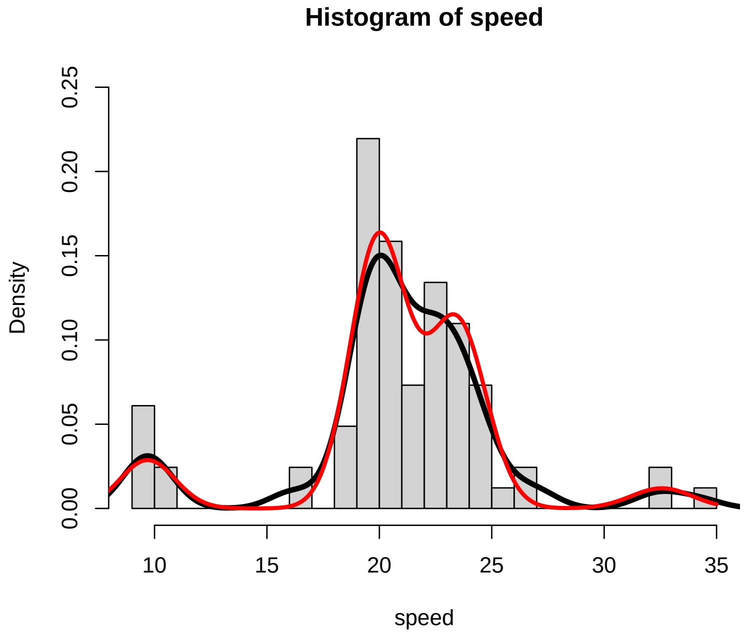


Figure 4.6: Density estimation for the variable speed in red with the Gibbs sampler, in black with the classic inference

4.3 Application of the Gibbs sampler to a real dataset

4.3.1 Concrete Compressive Data

The last dataset considered concerns concrete, more details can be find at the following page Gediya (2021). The Compressive Strength of Concrete determines the quality of Concrete. This is generally determined by a standard crushing test on a concrete cylinder. This requires engineers to build small concrete cylinders with different combinations of raw materials and test these cylinders for strength variations with a change in each raw material. One of the materials is *fly ash*. This is the established variable for tested the Gibbs sampler.

Unlike the datasets before considered, this is not thought for performing well, this represents real data. This means that estimate the density is more interesting, but also more difficult.

The chosen variable do not present an easily interpretable histogram, so using Mclust the number of components for the mixture model is set at six. The values for the parameters needed are $\nu^2 = 10$, $\alpha_j = 1700$, $\beta_j = 10$ and $\omega_j = 0.5$ for all $j = 1, \dots, 6$.

The result obtained is:

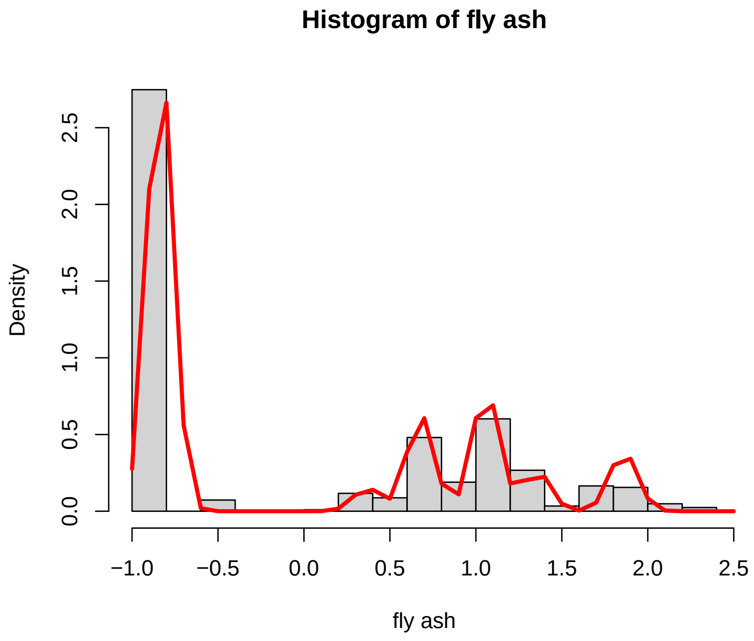


Figure 4.7: Density estimation for the variable fly ash

4.4 Observations after application

At the end of this chapter we would like to point out some results.

4.4.1 Label Switching and Burn-in period

Applying the algorithm allows us to observe occurrences such as label switching and the consequent need to set a burn-in period in order to eliminate iterations where convergence has not yet been achieved.

Consider once again the theoretical example described in the Section 4.1 with three components and assume that no burn-in period was provided. It is possible to see the trend of the values that the averages take as the iterations pass through a traceplot.

It is shown here:

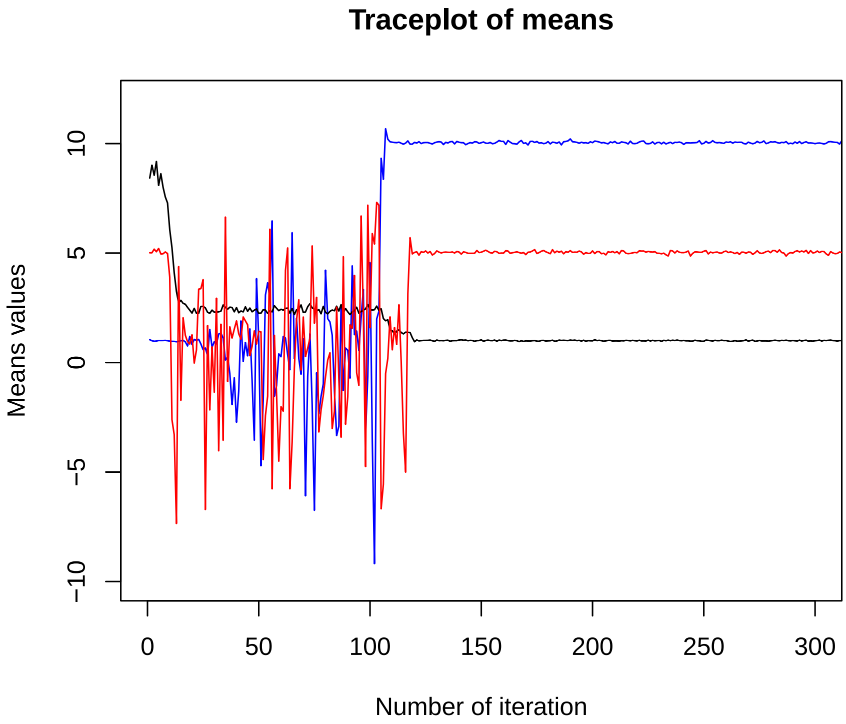


Figure 4.8: Traceplot of means

As can be observed, convergence is not immediately achieved. For the first 130 iterations, the averages continue to switch, until they reach stability around three distinct values. This phenomenon is the so-called *label switching* and the 130 iterations constitute the burn-in period. This number of iterations changes every time the algorithm is applied to the data. As consequence, a large number of iterations is usually chosen so that a high burn-in period can be set to guarantee, with certainty, that convergence can be achieved and still obtain acceptable results.

4.4.2 Starting values for the parameters

One of the delicate aspects of implementing the algorithm is the choice of starting values to be assigned to the parameters for the full conditionals. The Gibbs sampler is done in such a way that convergence is always guaranteed to be achieved. However, the choice of these initial parameters determines the speed with which convergence is reached and the quality of the density estimation.

One possible basic criterion for choosing these parameters is to look at the starting graph. The parameter β_j influences the smoothing of the distribution, whereas it follows from α_j that the greater the value taken, the more the peaks identified will tend to be accentuated. Lastly, ν^2 influences where the peaks will be centred.

Bibliography

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics 14, 1* , 1–13.

Cox, R. T. (1961). The algebra of probable inference. *The Johns Hopkins Press* .

Diaconis, P. & Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* 7 **(2)**, 269–281.

Gediya, V. (2021). Concrete Compressive Strength Data. https://www.kaggle.com/datasets/vivekgediya/concrete-data?resource=download.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2015). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Green, P. J. (2018). *Handbook of Mixture Analysis - Chapter 1*. Chapman and Hall/CRC.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer New York.

RDocumentation (2022a). RDocumentation. Galaxy Data. https://search.r-project.org/CRAN/refmans/bmixture/html/galaxy.html.

RDocumentation (2022b). RDocumentation. Old Faithful Geyser Data. https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/faithful.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer New York.

Robert, C. P. & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer New York.

Savage, L. J. (1954). *The foundations of statistics*. John Wiley & Sons Inc.

Savage, L. J. (1972). *The foundations of statistics, revised edn.* Dover Publications Inc.