The Islamic University of Gaza
Deanery of Higher Studies
Faculty of Science
Department of Mathematics

Bayesian Inference on Finite Mixtures of Exponential Distributions

Presented by
Ahlam M. Saleh

Supervisor
Prof. Mohamed I. Riffi

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master in Mathematics
2016

Abstract

Mixed distributions are widely used to model data in which each observation is assumed to come from one of a number of different groups. In this thesis, we investigate the Bayesian estimation for the finite exponential mixture model using the Gibbs sampler as an important one of the MCMC methods. Our approach in this thesis depends on using the Gibbs sampler to simulate a Markov chain which has the posterior density as its long-run (stationary) distribution. Then we use the resulting sample to make the suitable Bayesian computations and draw conclusion about the unknown parameters of the exponential mixture model. We conclude this thesis by presenting a real data example to illustrates our methodology.

Contents

List of Tables

List of Figures

Introduction

The main goal of this thesis is to use the Bayesian analysis to estimate the finite mixture of exponential distributions.

In some times the simple exponential distribution becomes unsuitable to model a data that contains a large amount of over dispersion. We face this challenge by using the exponential mixture model to describe the inhomogeneity within the population.

Mixture models are good alternative candidates to model data when simple models fail. In particular, finite mixture models can provide important information about the number of subpopulations include the entire population.

The Markov Chain Monte Carlo (MCMC) methods is a collection of tools that is one of the most important tools of the Bayesian statistical inference and computational statistics. MCMC is a class of methods in which we can simulate draws that are slightly dependent and are approximately from a (posterior) distribution. We then take those draws and calculate quantities of interest for the (posterior) distribution. In Bayesian statistics, there are generally two MCMC algorithms that we use: the Gibbs Sampler and the Metropolis-Hastings (M-H) algorithm.

The Gibbs sampler algorithm is one of the most basic Markov Chain Monte Carlo Methods that is used in Bayesian Analysis. It's used to draw samples from a distribution that is either hard to sample from or its probability density function (pdf) is only known up to a normalizing constant. The Gibbs sampler algorithm generates a Markov chain which has as its stationary distribution the posterior distribution by simulating observations from a different proposed distribution. This simulation procedure enables us to draw a sample from the posterior distribution that can be used in estimation and other statistical inference.

The most widely used finite mixture distributions are those involving normal components. Medgyessi (1961) analyzes absorption spectra in terms of

normal mixtures, to every theoretical "line" belongs an intensity distribution whose graph fits very well to that of some normal distributions, and also applies normal mixtures to the results of protein separation by electrophoresis. Bhattacharya (1967) studies the length distribution of a certain type of fish and finds it useful to split his observations into age categories, with each category contributing a normal component distribution to yield an overall mixture.

Clark et al. (1968) provides an illustration of an area in which mixture distributions are being applied more frequently namely the study of disease distributions.

This thesis is organized as follows:

Introduction

In the introduction, we briefly talk about mixture models and their importance. We also talk about the importance of the natural exponential mixtures in applications. Then we mention the approach we are going to follow in making Bayesian inference about the exponential mixtures.

Chapter 1 Introduction to Bayesian Statistics

This chapter includes the following topics. Bayes theorem, expressing the posterior probability density function in terms of the prior density and the likelihood function, conjugate priors, and some related examples. **Chapter 2 Finite Mixtures of Distributions**

We give in this chapter an introduction to finite mixtures models. Then, we present the finite exponential mixtures model using the missing data formulation.

Chapter 3 MCMC methods in Bayesian Inference

In this chapter, we give a brief introduction on discrete-time Markov chains and the Gibbs sampler and algorithm as one of the most basic Markov Chain Monte Carlo (MCMC) methods in Bayesian analysis. We present the algorithm used to generate samples.

Chapter 4 Bayesian Analysis of Finite Exponential Mixtures

We use in this chapter the Gibbs sampler and algorithm to draw samples from the posterior of the exponential mixtures in order to use them in the Bayesian analysis. This can be done by using the R language. We use these samples in the estimation of the unknown parameters of the model.

Chapter 1

Introduction to Bayesian Statistics

The possibilities are commonly used in our daily lives, 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 this chapter we will introduce a mathematical methods to find probability, it can be shown that probabilities can numerically represent a set of rational beliefs, that there is a relationship between probability and information, and that Bayes rule provides a rational method for updating beliefs in light of new information. The process of inductive learning via Bayes rule is referred to as Bayesian inference.

1.1 Introduction

Bayesian statistics is based on the theorem first discovered by Reverend Thomas Bayes and published after his death in the paper " An Essay Towards Solving a Problem in the Doctrine of Chances " by his friend Richard Price.

Bayes' theorem combines the two sources of information about the unknown parameter value: the prior density and the observed data. The prior density gives our relative belief weights of every possible parameter value before we observe the data. The likelihood function gives the relative weights to every possible parameter value that comes from the observed data. Bayes' theorem combines these into the posterior density, which gives our relative belief weights of the parameter value after observing the data. See [2]

In this section, we introduce the basics of bayesian statistics and we will explain how bayesian statistic different with other non bayesian models. The Frequentist and Bayesian approaches to statistics differ in the definition of probability. For a Frequentist, probability is the relative frequency of the occurrence of an event in a large set of repetitions of the experiment. In Bayesian statistics, on the other hand, probability is not defined as a frequency of occurrence but as the plausibility that a proposition is true, given the available information. Bayesian statistical analysis is concerned with calculating probability distributions of parameters in statistical models, where this statistical model describes the relationship between the parameters and the data in a mathematical model. Bayesian statistical analysis treats the parameters as random variables. A non-Bayesian statistical analysis treats parameters as fixed values without distribution. so we can summarize the differs between bayesian inference and classical, frequentist inference in four ways:

1. Frequentist inference estimates the probability of the data having occurred given a particular hypothesis ($P(Y|H)$) whereas Bayesian inference provides a quantitative measure of the probability of a hypothesis being true in light of the available data ($P(H|Y)$);

2. Their definitions of probability differ: frequentist inference defines probability in terms of infinite relative frequencies of events, whereas Bayesian

inference defines probability as a degree of belief in the likelihood of an event.

3. Bayesian inference uses prior knowledge along with the sample data whereas frequentist inference uses only the sample data;

4. Bayesian inference treats model parameters as random variables whereas frequentist inference considers them to be estimates of fixed, true quantities.

Components of bayesian statistical analysis:

- The prior distribution: which the probability of observing the parameter that is expected by the investigator before the experiment is conducted.

- Likelihood distribution: based on modeling assumptions, how [relatively] likely the data Y are if the truth is β, denoted $f\left(Y|\beta\right)$

- posterior distribution: stating what we know about the parameter β, combining the prior with the data Y denoted $P\left(\theta|y\right)$

Computational Bayesian Statistics

The main ideas of computational Bayesian statistics is finding the posterior distribution using bayes theorm so, we will introduce some rules to explain bayes theorm:

Definition 1.1.1. [9](**Probability Axioms**) Let S be the sample space of a random phenomenon. Suppose that to each event A of S, a number denoted by $P(A)$ is associated with A. If P satisfies the following axioms, then it is called a **probability** and the number $P(A)$ is said to be the **probability** of A.

Axiom 1 $P\left(A\right) \geq 0$.

Axiom 2 $P(S) = 1$.

Axiom 3 If $\{A_1, A_2, A_3, ...\}$ is a sequence of mutually exclusive events (i.e., the joint occurrence of every pair of them is impossible: $A_i \cap A_j = \phi$ when $i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Definition 1.1.2. [9](**Conditional Probability**). Let A and B be two events with $P(A) > 0$ and $P(B) > 0$. Then the conditional probability of A given B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Definition 1.1.3. [14]**(Partition)** A collection of sets $\{H_1, ..., H_k\}$ is a partition of another set $\mathcal{H}$ if

1. the events are disjoint, which we write as $H_i \cap H_j = \phi$; for $i \neq j$;
2. the union of the sets is $\mathcal{H}$, which we write as $\cup_{i=1}^{k} H_i = \mathcal{H}$.

Partitions and probability

Suppose $\{H_1, ..., H_k\}$ is a partition of $\mathcal{H}$, $P(H) = 1$, and E is some specific event. The axioms of probability imply the following:

Rule of total probability: $\sum_{i=1}^{k} P(H_i) = 1$

Rule of marginal probability:

$$\begin{aligned} P(E) &= \sum_{i=1}^{k} P(E \cap H_i) \\ &= \sum_{i=1}^{k} P(E|H_i) P(H_i) \end{aligned}$$

Theorem 1.1.4. *[9]**(Law of total probability)** If $\{B_1, B_2, ..., B_n\}$ is a partition of the sample space S of an experiment and $P(B_i) > 0$ for $i =$*

$1, 2, ..., n$, *then for any event* A *of* S,

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... + P(A|B_n)P(B_n) \\ &= \sum_{i=1}^{n} P\left(A|B_i\right) P\left(B_i\right) \end{aligned}$$

More generally, let $\{B_1, B_2, ...\}$ *be a sequence of mutually exclusive events of* S *such that* $\bigcup_{i=1}^{\infty} B_i = S$. *Suppose that, for all* $i \geq 1$, $P(B_i) > 0$. *Then for any event* A *of* S,

$$P(A) = \sum_{i=1}^{\infty} P\left(A|B_i\right) P\left(B_i\right)$$

Bayes rule for discrete random variables:

$$\begin{aligned} P\left(H_j|E\right) &= \frac{P\left(E|H_j\right) P\left(H_j\right)}{P(E)} \\ &= \frac{P\left(E|H_j\right) P(H_j)}{\sum_{i=1}^{k} P\left(E|H_i\right) P(H_i)} \end{aligned}$$

provided $P(E) > 0$. In this expression of Bayes Theorem, $P(H|E)$ is the probability of H after obtaining E, and $P(H)$ is the prior probability of H before considering E. The conditional probability on the left-hand side of the theorem, $P(H|E)$, is usually referred to as the posterior probability of H.

Bayes' rule for continous random variables:

let the unknown parameter be θ, and denote the data available for analysis as $Y = (y_1, y_2, ..., y_n)$. In the case of continuous parameters, beliefs about the parameter are represented as probability density functions or pdfs; we denote the prior pdf as $P(\theta)$ and the posterior pdf as $P(\theta|y)$.

Then, Bayes Theorem for a continuous parameter is as follows:

$$P\left(\theta|Y\right) = \frac{P\left(Y|\theta\right) P(\theta)}{\int P\left(Y|\theta\right) P(\theta) d\theta} \tag{1.1}$$

This distribution is called the **posterior distribution**. The denominator of the above equation is called the **normalizing constant**. The Bayesian

inference proceeds from the posterior distribution.
Let

$$z = \int_{-\infty}^{\infty} P\left(Y|\theta\right) P(\theta) d\theta$$

Equation (2.1) becomes:

$$\overbrace{P\left(\theta|Y\right)}^{posterior} = \frac{\overbrace{P\left(Y|\theta\right)}^{likelihood}\overbrace{P(\theta)}^{prior}}{\underbrace{z}_{normalizingconstant}} \tag{1.2}$$

$$\propto P(Y|\theta)P(\theta) \tag{1.3}$$

in words:

$$posterior \propto likelihood \times prior$$

1.2 Likelihood Function

In statistics, a likelihood function (often simply the likelihood) is a function of the parameters of a statistical model. Likelihood functions play a key role in statistical inference, especially methods of estimating a parameter from a set of statistics.

Definition 1.2.1. The likelihood of a set of parameter values, θ, given outcomes x, is equal to the probability of those observed outcomes given those parameter values, that is

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

The likelihood function is defined differently for discrete and continuous probability distributions.

The likelihood for discrete distribution

let X be a random variable with a discrete probability distribution p depending on a parameter θ. Then the function

$$\mathcal{L}(\theta|x) = P_\theta(X = x)$$

considered as a function of θ, is called the likelihood function (of θ, given the outcome x of X). Sometimes the probability on the value x of X for the parameter value θ is written as $P(X = x|\theta)$; often written as $P(X = x; \theta)$ to emphasize that this value is not a conditional probability, because θ is a parameter and not a random variable.

The likelihood for continuous distribution

Let X be a random variable with a continuous probability distribution with density function f depending on a parameter θ. Then the function

$$\mathcal{L}(\theta|x) = f_\theta(x),$$

considered as a function of θ, is called the likelihood function (of θ, given the outcome x of X). Sometimes the density function for the value x of X for the parameter value θ is written as $f(x|\theta)$, but should not be considered as a conditional probability density.

Definition 1.2.2. The Joint Likelihood Function: The likelihood function of the random sample is the product of the individual observation likelihoods.

Example 1.2.3. [2] Suppose $y_1, y_2, ..., y_n$ are a random sample from a normal (μ, σ^2) distribution where the variance σ^2 is a known constant. For a random sample, the joint likelihood is the product of the individual likelihoods, so it

is given by

$$\begin{aligned} f(y_1, ..., y_n|\mu) &= f(y_1|\mu) \times f(y_2|\mu) \times ... \times f(y_n|\mu) \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_1-\mu)^2/2\sigma^2} \times ... \times \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_n-\mu)^2/2\sigma^2} \\ &\propto \prod_{i=1}^{n} e^{-(y_i-\mu)^2/2\sigma^2} \\ &\propto e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2} \end{aligned}$$

Multiply out the terms in the exponent and collect like terms, and using the rule that

$$\bar{y} = \frac{\sum y_i}{n} \Longrightarrow \sum y_i = n\bar{y}$$

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \mu)^2 &= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} 2\mu y_i + \sum_{i=1}^{n} \mu^2 \\ &= n\mu^2 - 2\mu n\bar{y} + \sum_{i=1}^{n} y_i^2 \end{aligned}$$

Factor out n from the first terms and complete the square

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \mu)^2 &= n(\mu^2 - 2\mu\bar{y}) + \sum_{i=1}^{n} y_i^2 \\ &= n(\mu - \bar{y})^2 - n(\bar{y})^2 + \sum_{i=1}^{n} y_i^2 \end{aligned}$$

Put this back in the likelihood, and absorb the part that does not affect the shape into the constant

$$\begin{aligned} f\left(y_1, ..., y_n|\mu\right) &\propto e^{-\frac{1}{2\sigma^2/n}}(\mu - \bar{y})^2 \times e^{-\frac{1}{2\sigma^2}[n(\bar{y}^2) + \sum_{i=1}^{n} y_i^2]} \\ &\propto e^{-\frac{1}{2\sigma^2/n}(\mu-\bar{y})^2} \end{aligned}$$

We recognize this is the likelihood of y. It is a normal distribution with mean μ and variance $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$. Thus the likelihood of the whole random sample

is proportional to the likelihood of $\bar{y}$, a single draw from a normal $\left(\mu, \sigma_{\bar{y}}^2\right)$, where $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$.

Example 1.2.4. [10] Let $X_1, X_2, ..., X_n$ be an iid exponential (θ).
Suppose the prior density is given by:

$$f(\theta) = e^{-\theta}, \quad \theta > 0.$$

The likelihood density is:

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} (\theta e^{-\theta x_i}) \\ &= (\theta e^{-\theta x_1})(\theta e^{-\theta x_2})...(\theta e^{-\theta x_n}) \\ &= \underbrace{\theta.\theta.....\theta}_{n-times}(e^{-\theta x_1}.e^{-\theta x_2}.....e^{-\theta x_n}) \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \end{aligned}$$

The posterior density is:

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta) f(\theta) \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} x_i}.e^{-\theta} \\ &= \theta^n e^{-\theta(1+\sum_{i=1}^{n} x_i)} \end{aligned}$$

Clearly, this is the density of gamma distribution with parameters $n+1$ and $1 + \sum_{i=1}^{n} x_i$.
So,

$$(\theta|x) \sim \text{gamma}(n+1, 1 + \sum_{i=1}^{n} x_i).$$

Table 1. Notation for common pdf's and pmf's

Table 1.1: pdf,s or pmf

Name	pdf or pmf	parameter(s)
Beta	$Be(\alpha,\beta) = \frac{1}{beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, $x \in (0,1)$.	$\alpha > 0$, $\beta > 0$.
Binomial	$Bi(x\|n,\theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x(1-\theta)^{n-x}$, $x \in \{0,1,...,n\}$.	$n \in \{1,2,...\}$, $\theta \in (0,1)$.
Exponential	$Ex(x\|\theta) = \theta e^{-\theta x}$, $x>0$	$\theta > 0$.
Gamma	$Ga(x\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$, $x>0$	$\alpha > 0$, $\beta > 0$.
NegBinomial	$Nb(x\|r,\theta) = \begin{pmatrix} r+x-1 \\ r-1 \end{pmatrix} \theta^r(1-\theta)^x$, $x \in \{1,2,...\}$.	$r \in \{1,2,...\}$, $\theta \in (0,1)$.
Normal	$N(x\|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in R$.	$\mu \in R$, $\sigma > 0$.
Poisson	$Pn(x\|\lambda) = \frac{1}{x!}e^{-\lambda}\lambda^x$, $x \in \{0,1,...,n\}$	$\lambda > 0$.
Inverse Gamma	$IGa(x\|\alpha,\beta) = \frac{\Gamma(\alpha)}{\beta^\alpha}\frac{1}{x^{\alpha-1}}e^{-\frac{\beta}{x}}$, $x>0$	$\alpha > 0$, $\beta > 0$.

1.3 Prior Distribution

In Bayesian statistical inference, a prior probability distribution, often called simply the prior, of an uncertain quantity is the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account. $f(\theta)$ is the prior probability of obtaining the specified parameter. In other words, $P(\theta)$ is the probability of observing θ that is expected by the investigator befor the experiment is conducted, a prior can be the purely subjective assessment of an experienced expert. See[36]

In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. When a family of conjugate priors exists, choosing a prior from that family simplifies calculation of the posterior distribution, so we will introduce the definition of conjugete prior. See[38]

Definition 1.3.1. [38] Suppose a prior density $p(\theta)$ belongs to a class of parametric of densities, $\mathcal{F}$. Then the prior density is said to be conjugate with respect to a likelihood $p(y|\theta)$ if the posterior density $p(\theta|y)$ is also in $\mathcal{F}$.

The conjugate family. The conjugate family of priors for a member of the one-dimensional exponential family of densities has the same form as the likelihood. It is given by

$$g(\theta) \propto A(\theta)^k e^{C(\theta)\times l} \tag{1.4}$$

where k and l are the constants that determine its shape. See[2]
The following table conclude some distributions and their conjugate prior.

Table 1.2: conjugate prior for some distribution

Likelihood	Model parameters	Conjugate prior distribution	Prior parameters
Binomial	p (probability)	Beta	α , β
Poisson	λ (rate)	Gamma	K , θ
Geometric	p (probability)	Beta	α , β
Negative binomial	p (probability)	Beta	α , β
Exponential	λ (rate)	Gamma	α , β
Gamma	β (rate)	Gamma	α , β

Example 1.3.2. Conjugate prior for exponential(λ) is gamma(α, β). The conjugate prior for λ will have the same form as the likelihood. we will choose$A(\lambda)$, $C(\lambda)$ in 1.4 as follow
$A(\lambda) = \lambda$, $C(\lambda) = -\lambda$
Thus its shape is given by

$$g(\lambda) \propto \lambda^k e^{-\lambda l}$$

We recognize this to be the gamma(α, β) distribution which has exact density given by

$$g(\lambda) = \frac{\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta\lambda},$$

where $\alpha - 1 = k$ and $\beta = l$.

1.4 Posterior Probability

In some cases we have a formula for the exact posterior. In other cases we only know the shape of the posterior using Bayes' theorem. The posterior is proportional to prior times likelihood. In those cases we can find the posterior density numerically by dividing through by the scale factor needed to make the integral of the posterior over its whole range of values equal to one. This scale factor is found by integrating the prior times likelihood over the whole

range of parameter values. Thus the posterior is given by

$$g(\theta|y_1, y_2, ..., y_n) = \frac{P(y_1, y_2, ..., y_n|\theta)P(\theta)}{\int P(y_1, y_2, ..., y_n|\theta)P(\theta)d\theta} \tag{1.5}$$

Definition 1.4.1. The posterior probability is the probability of the parameter θ given the evidence X : $p(\theta|X)$.

It contrasts with the likelihood function, which is the probability of the evidence given the parameters: $p(X|\theta)$.

The two are related as follows:

Let us have a prior belief that the probability distribution function is $p(\theta)$ and observations x with the likelihood $p(x|\theta)$, then the posterior probability is defined as

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

The posterior probability can be written in the memorable form as

Posterior probability $\propto$ Likelihood $\times$ Prior probability

Example 1.4.2. (continued) If t is the waiting time from the exponential(λ) distribution and we use the gamma(α, β) prior distribution for λ as in example (2.3.2), the shape of the posterior is given by:

$$Posterior \propto Likelihood \times Prior$$

$$\begin{aligned} g(\lambda|t) &\propto \frac{\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)}e^{-\beta\lambda} \times \lambda e^{-\lambda t} \\ &\propto \lambda^{\alpha-1+1}e^{-(\beta+t)\lambda} \\ &\propto \lambda^{\grave{\alpha}-1}e^{\grave{\beta}\lambda} \end{aligned}$$

where the constants are updated by $\grave{\alpha} = \alpha + 1$ and $\grave{\beta} = \beta + t$. We recognize this to be the gamma ($\grave{\alpha}$,$\grave{\beta}$) distribution.

Example 1.4.3. [10] Let $X_1, X_2, ..., X_n$ be an i.i.d sample from the exponential distribution with density

$$f(x|\lambda) = \lambda e^{-\lambda x}, \qquad x > 0, \quad \lambda > 0.$$

Suppose the prior density for is given by:

$$f(\lambda) = \mu e^{-\mu\lambda}, \qquad \lambda > 0, \quad for \quad some \quad known \quad \mu > 0.$$

The likelihood density is:

$$\begin{aligned} f(x|\lambda) &= \prod_{i=1}^{n} f(x_i|\lambda) \\ &= \prod_{i=1}^{n} (\lambda e^{-\lambda x_i}) \\ &= (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})...(\lambda e^{-\lambda x_n}) \\ &= \underbrace{\lambda.\lambda.....\lambda}_{n-copies} e^{-\lambda x_1 - \lambda x_2 - ... - \lambda x_n} \\ &= \lambda^n e^{-\lambda(x_1+x_2+...+x_n)} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

Therefore,

$$f(x|\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

The posterior density is:

$$\begin{aligned} f(\lambda|x) &\propto f(x|\lambda) f(\lambda) \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \mu e^{-\mu\lambda} \\ &= \mu \lambda^n e^{-\mu\lambda - \lambda \sum_{i=1}^{n} x_i} \\ &= \mu \lambda^n e^{-(\mu + \sum_{i=1}^{n} x_i)\lambda} \end{aligned}$$

We do not write the term μ which does not involve λ.
The posterior density becomes:

$$f(\lambda|x) \propto \lambda^n e^{-(\mu + \sum_{i=1}^{n} x_i)\lambda}$$

Clearly this is the density of a gamma distribution with parameters:

$$n+1 \quad and \quad \sum_{i=1}^{n} x_i + \mu$$

Therefore,

$$f(\lambda|x) = Ga(n+1, \sum_{i=1}^{n} x_i + \mu).$$

1.4.1 Main Points In Bayesian Analysis

- Bayesian statistics does inference using the rules of probability directly.
- Bayesian statistics based on Bayes theorem which is the basic tool we use it to find posterior distribution of parameters, and this theorem find the posterior by the combining with prior $g(\theta_1, ..., \theta_k)$ and likelihood $f(y_1, ..., y_n, \theta_1, ..., \theta_k)$ as explained before.
- It is easy to find the missing posterior by posterior proportional to prior times likelihood and in symbols:

  $$g(\theta_1, ..., \theta_k|y_1, ..., y_n) \propto g(\theta_1, ..., \theta_k) \times f(y_1, ..., y_n, \theta_1, ..., \theta_k)$$

  and this posterior is not the exact posterior density, it must be divided by its integral to make it exact.
- Evaluating the integral may be very difficult, particularly if there are lots of parameters. It is hard to find the exact posterior except in a few special cases.
- In view of the previous two points, then if two experiments have proportional likelihoods, then they should lead to the same inference. See[2]

Chapter 2

Finite Mixtures of Distributions

2.1 Introduction to Finite Mixture Models

Finite mixture of distributions have a great importance in statistical modeling of a wide variety of random phenomena. In the past decade the extent and the potential of the applications of finite mixture models have widened considerably. Because of their flexibility mixture models are being increasingly exploited as convenient way in which to model unknown distributional shapes. The mixture models are useful when we need to divide the main population into subpopulations. For example, the distribution of heights in a population of adults reflects the mixture of males and females in the population. so in this case mixtures is best to model male and female heights as separate univariate, perhaps normal, distributions, rather than a single bimodel distribution. This make us to introduce the following definition.

Definition 2.1.1. Let $\mathcal{F}$ be a family of distribution functions. A random variable X is said to have a finite mixture distribution if its distribution function F satisfies

$$F(x) = \sum_{j=1}^{k} \pi_j F_j(x)$$

where the $F_j \in \mathcal{F}$ are the distinct distribution functions of the k mixture components or populations, and the mixture proportions satisfy $\pi_j > 0$, $j = 1, ..., k$ and $\sum_{j=1}^{k} \pi_j = 1$,

but in this thesis we consider models in which data $x^n = x_1, ..., x_n$, are assumed to be independent observations from a mixture density with k (k possibly unknown but finite) components:

$$p(x|\pi, \phi, \eta) = \pi_1 f(x; \phi_1, \eta) + ... + \pi_k f(x; \phi_k, \eta) \tag{2.1}$$

where $\pi = (\pi_1, ..., \pi_k)$ are the *mixture proportions* which are constrained to be non-negative and sum to unity; $\phi = (\phi_1, ..., \phi_k)$ are the (possibly vector) component specific parameters, with ϕ_i being specific to component i; and η is a (possibly vector) common parameter which is common to all components. Throughout this thesis $p(.|.)$ will be used to denote conditional densities and distributions.

Remark 2.1.2. [31] The finite mixture represented by $f(x) = \sum_{j=1}^{k} p_j f(x|\theta_j)$ is said to be identifiable if we have two representations

$$f(x) = \sum_{j=1}^{k} p_j f(x|\theta_j) \quad \text{and} \quad f^*(x) = \sum_{j=1}^{k^*} p_j^* f(x|\theta_j^*),$$

then $f \equiv f^*$ if and only if $k = k^*$ and there exists a permutation π of the indexes $(1, \ldots, k)$ such that $p_j = p_{\pi_j}^*$ and $\theta_j = \theta_{\pi_j}^*$.

2.1.1 Missing data

Mixture distributions are typical examples of missing data models. In mixture model we divide data into subgroups and each subgroup of the data

is called cluster. In the case that the clusters are known, the problem of estimating the parameters becomes so much simpler. It is convenient to introduce the missing data formulation of the model, in which each observation x_j is assumed to arise from specific but unknown (that is missing) component z_j of the mixture. Here z_n is multinomial random variable and

$$z_n^k = \begin{cases} 1 & \text{if instance } n \text{ is from component } k \text{ ,} \\ 0 & \text{otherwise.} \end{cases}$$

We will refer to the missing data $z^n = z_1, ..., z_n$ as allocation variables, and to (x^n, z^n) as the completed data. the model 2.1 can be written in terms of the missing data, with $z_1, ..., z_n$ assumed to be relations of independent and identically distributed discrete random variables $Z_1, ..., Z_n$ with probability mass function

$$P(Z_j = i|\pi, \phi, \eta) = \pi_i, \qquad j = 1, ..., n; \quad i = 1, ..., k$$

Conditional on the Zs, $x_1, ..., x_n$ are assumed to be independent observations from the densities

$$p(x_j|Z_j = i, \pi, \phi, \eta) = f(x_j; \phi_i, \eta), \quad j = 1, ..., n.$$

Integrating out the missing data $Z_1, ..., Z_n$ then yields the model 2.1:

$$p(x_j|\pi, \phi, \eta) = \sum_{i=1}^{k} P(Z_j = i|\pi, \phi, \eta) p(x_j|Z_j = i, \pi, \phi, \eta)$$

$$= \sum_{i=1}^{k} \pi_j f(x_j; \phi_i, \eta).$$

Now to show why we use missing data model:

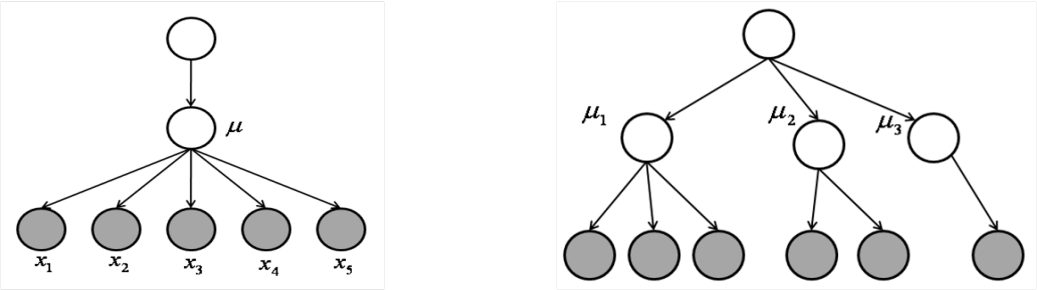

Figure 2.1: Graphical models

Figure (a) represents the typical case when the population have the same mean, but in Figure (b) the population is not homogenous so, we divide it into subpopulations with different means, this case is a good model for non-homogenous data, this type of models are called mixture models. Each group of the data are called cluster (In this case we have $k = 3$ clusters)
Now, if the mixture distributions π_k are given and the necessary parameters for each component θ_k, we can write the likelihood as

$$p(x, z|\theta, \pi) = \prod_n \sum_k z_n^k p(x_n|\theta_k)\pi_k \tag{2.2}$$

where for each k, only one of the terms z_n^k is equal to 1 and the rest are zero. We call this the **complete data likelihood**. The full computations of the corresponding posterior distribution involves the expansion of the likelihood

$$p(x, z|\theta, \pi) = \prod_n \sum_k z_n^k p(x_n|\theta_k)\pi_k$$

into a sum of K^n terms, and this computationally too expensive to be used, and this difficulty explain why we use the missing data model, so we introduce the following example to explain the meaning:

Example 2.1.3. In order to show how to derive the complete data likelihood, we consider the following example. We consider data coming from two clusters ($K = 2$) with four data points ($N = 4$). We denote the parameters for clusters by λ_1 and λ_2. the data are assumed to come from exponential distribution which has the following probability density function

$$p(x|\lambda) = \lambda e^{-\lambda x}, 0 < x < \infty$$

The data points and their clusters are given in the following table

	x_i	K
1	2	1
2	7	2
3	3	1
4	9	2

The complete data likelihood equals

$$\begin{aligned}
p(x, z|\lambda, \pi) &= \prod_n \sum_k z_n^k p(x_n|\lambda_k)\pi_k \\
&= \prod_n \sum_k z_n^k (\lambda_k e^{-\lambda_k x_n})\pi_k \\
&= (z_1^1 \lambda_1 e^{-2\lambda_1}\pi_1 + 0)(0 + z_2^2 \lambda_2 e^{-7\lambda_2}\pi_2)(z_3^1 \lambda_1 e^{-3\lambda_1}\pi_1 + 0)(0 + z_4^2 \lambda_2 e^{-9\lambda_2}\pi_2) \\
z_1^1 = z_2^2 = z_3^1 = z_4^2 = 1, \quad & z_1^2 = z_2^1 = z_3^2 = z_4^1 = 0 \\
&= (\lambda_1 e^{-2\lambda_1}\pi_1)(\lambda_2 e^{-7\lambda_2}\pi_2)(\lambda_1 e^{-3\lambda_1}\pi_1)(\lambda_2 e^{-9\lambda_2}\pi_2) \\
&= (\pi_1(\lambda_1 e^{-x_1\lambda_1}))^1 (\pi_2(\lambda_2 e^{-x_1}\lambda_2))^0 (\pi_1(\lambda_1 e^{-x_2\lambda_1}))^0 (\pi_2(\lambda_2 e^{-x_2\lambda_2}))^1 \\
&(\pi_1(\lambda_1 e^{-x_3\lambda_1}))^1 (\pi_2(\lambda_2 e^{-x_3\lambda_2}))^0 (\pi_1(\lambda_1 e^{-x_4\lambda_1}))^0 (\pi_2(\lambda_2 e^{-x_4\lambda_2}))^1 \\
&= \prod_{i=1}^n \prod_{j=1}^k z_{ij} p_j f(x_i|\theta_j) \\
&= \prod_{i=1}^n \prod_{j=1}^k (p_j f(x_i|\lambda_j))^{z_{ij}}
\end{aligned}$$

Remark 2.1.4. The original definition of mixture distribution is given in equation 2.2, but after using the missing data formulation we invert the K^n terms in the summation into one term when $i = j$, so we replace summation in 2.2 by the product over the values of K, so we get the new expression of the definition of the mixture which given by:

$$f(x, z|\lambda, p) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j f(x_i|\lambda_j))^{z_{ij}} \tag{2.3}$$

2.2 Exponential distribution

On of the most important continous distributions is the exponential distribution, exponential distribution is important in the analysis of failure data, where the probability density function of failure time can be approximated by :

$$f(x) = \frac{1}{\mu} exp(\frac{-x}{\mu}), x \geq 0; \mu > 0$$

Thus, exponential distribution can be used in the following:

- The mean time between failures of a machine in a factory.
- The length of time one can expect a complex system to function without failing.
- The probability density function for the times of scores in game may be exponentially distributed.
- The interarrival time between two customers at a post office
- The time between two accidents at an intersection
- The time until the next baby is born in a hospital
- The time interval between the observation of two consecutive shooting stars on a summer evening
- The time between two consecutive fish caught by a fisherman from a large lake with lots of fish. See[9]

Definition 2.2.1. [6] A continuous random variable X is called **exponential** with parameter $\lambda > 0$ if its density function is given by:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if t} \geq 0 \ , \\ 0 & \text{if t} < 0. \end{cases}$$

and for all n ≥ 1 :

$$P(X_n \leq t) = F(t) \begin{cases} 1 - e^{-\lambda t} & \text{if t} \geq 0 \ , \\ 0 & \text{if t} < 0. \end{cases}$$

Remark 2.2.2. In the exponential distribution λ is the average number of the events in one time unit.

let X be an exponential random variable with parameter λ then:

$$E(X) = \int_{-\infty}^{\infty} xf(x)d(x) = \int_{0}^{\infty} x(\lambda e^{-x\lambda})d(x) = \frac{1}{\lambda}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)d(x) = \int_{0}^{\infty} x^2(\lambda e^{-x\lambda})d(x) = \frac{2}{\lambda^2}$$

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Example 2.2.3. [9] Suppose that every three months , on average, an earthquake occurs in California. What is the probability that the next earthquake occurs after three but before seven months?

Solution: Let X be the time (in months) until the next earthquake; it can be assumed that X is an exponential random variable with $\frac{1}{\lambda} = 3$ or $\lambda = \frac{1}{3}$. Now, to calculate $P(3 < X < 7)$, note that since F the distribution function of X, is given by:

$$F(T) = P(X \leq t) = 1 - e^{-t/3}, for t > 0$$

We can write

$P(3 < X < 7) = F(7)$ - $F(3) = (1 - e^{-7/3})$ - $(1 - e^{-1}) \approx 0.27$

Memory property:[6] It is traditional to formulate this property in terms of waiting for an unreliable bus driver. In words, if weve been waiting for t units of time then the probability we must wait s more units of time is the same as if we havent waited at all. In symbols

$$P(X > s + t | X > t) = P(X > s) \tag{2.4}$$

An important property of exponential distribution is memoryless property, to show that exponential distribution has this property recall that if $B \subset A$, then $P(B|A)$=$P(B)/P(A)$, so

$$P(X > s+t|X > t) = \frac{P(X > s+t)}{P(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s) \quad (2.5)$$

Example 2.2.4. [9] The lifetime of a TV tube (in years) is an exponential random variable, with mean 10. If Jim bought his TV set 10 years ago, what is the probability that its tube will last another 10 years? solution: let X be the lifetime of the tube. Since X is an exponential random variable, there is no deterioration with age of the tube. Hence
$P(X > 20|X > 10) = P(X > 10) = 1 - P(X \leq 10) = 1 - [1 - e^{-(1/10)10}] \approx 0.37$

2.3 Finite Exponential Mixtures Model

One of the most important as a component of a mixture is the exponential distribution, in this section we will present the finite exponential mixtures model using the missing data formulation.
The general form of a finite exponential mixture is

$$\begin{aligned} f(x) &= \sum_{i=1}^{k} \pi_i \frac{1}{\mu_i} exp(\frac{-x}{\mu_i}), \qquad x \geq 0 \\ &= 0 \qquad otherwise \end{aligned}$$

Chapter 3

MCMC Methods in Bayesian Inference

In this chapter, we will introduce the definition of the discrete-time Markov chains and the Gibbs sampler and algorithm as one of the most basic Markov Chain Monte Carlo (MCMC) methods in Bayesian analysis. We present the algorithm used to generate samples.

3.1 Markov Chains

In this section we will introduce the definition of discrete-time Markov chains and some examples which explain the meaning. We will discuss some basic properties of a Markov chain, basic concepts and notations. Some important theorems will be discussed also.

A Markov chain is a stochastic process with finite state space and Markov property which refers to the memoryless property of a stochastic process. A stochastic process has the Markov property if the conditional probability distribution of future states of the process depends only upon the present state, not on the sequence of events that preceded it. A process with this

property is called a Markov process.
Suppose that the random variables $X_0, X_1, ..., X_n$ represent the outcomes of some random experiment, and a finite Markov chain is a process which moves among the elements of a finite set S, where S is a state space for this system which contains the outcomes, so $S = \{x_0, x_1, ..., x_n\}$.
We denote by $p(i, j)$ the probability of moving from the state i to the state j in one step.

Example 3.1.1. [11] Let us begin with a simple example. We consider a random walker in a very small town consisting of four streets, and four street-corners v_1, v_2, v_3, and v_4 arranged as in Figure 3.1.1. At time 0, the random walker stands in corner v_1. At time 1, he flips a fair coin and moves immediately to v_2 or v_4 according to whether the coin comes up heads or tails. At time 2, he flips the coin again to decide which of the two adjacent corners to move to, with the decision rule that if the coin comes up heads, then he moves one step clockwise in Figure 3.1.1, while if it comes up tails, he moves one step counterclockwise. This procedure is then iterated at times 3, 4, ...
For each n, let X_n denote the index of the street-corner at which the walker stands at time n. Hence, $(X_0, X_1, ...)$ is a random process taking values in $\{1, 2, 3, 4\}$. Since the walker starts at time 0 in v_1, we have,

$$P(X_0 = 1) = 1. \tag{3.1}$$

Next, he will move to v_2 or v_4 with probability $1/2$ each, so that

$$P(X_1 = 2) = 1/2 \tag{3.2}$$

and,

$$P(X_1 = 4) = 1/2. \tag{3.3}$$

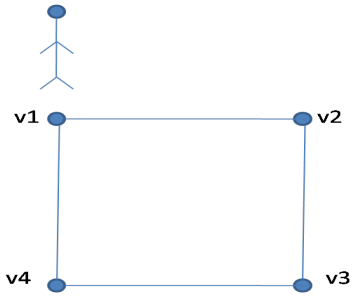

Figure 3.1: A random walker in a very small town.

To compute the distribution of X_n for n $\geq$ 2, Suppose that at time n, the walker stands at, say, v_2. Then we get the conditional probabilities

$$P(X_{n+1} = v_1 | X_n = v_2) = 1/2$$

and,

$$P(X_{n+1} = v_3 | X_n = v_2) = 1/2,$$

because of the coin-flipping mechanism for deciding where to go next. In fact, we get the same conditional probabilities of the process up to time n, i.e.,

$$P(X_{n+1} = v_1 | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}, X_n = v_2) = 1/2$$

and

$$P(X_{n+1} = v_3 | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}, X_n = v_2) = 1/2$$

for any choice of $i_0, ..., i_{n-1}$. (This is because the coin flip at time $n+1$ is independent of all previous coin flips, and hence also independent of $X_0, ..., X_n$.) This phenomenon is called the **memoryless property**, also known as the **Markov property**: the conditional distribution of X_{n+1} given $(X_0, ..., X_n)$ depends only on X_n.

Remark 3.1.2. The conditional distribution of X_{n+1} given that $X_n = v_2$ (say) is the same for all n. (This is because the mechanism that the walker uses to decide where to go next is the same at all times.), and this property is known as **time homogeneity**, or simply **homogeneity**.

We introduce this example for a general definition:

Definition 3.1.3. [9] A stochastic process $\{X_n : n = 1, 2, ...\}$ with finite or countably infinite state space S is said to be **Markov chain**, if for all $i, j, i_0, ..., i_{n-1} \in S$, and $n = 0, 1, 2, ...,$

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{n+1} = j | X_n = i). \tag{3.4}$$

Remarks 3.1.4.

- The elements of the state space S are not necessarily nonnegative integers (or numbers). But it is common to label the elements of S by nonnegative integers.
- If S is finite, the Markov chain is called **a finite Markov chain** or **a finite-state Markov chain.**
- If S is infinite, it is called **an infinite Markov chain** or **an infinite-state Markov chain.**
- The main property of a Markov chain, expressed by 3.4, is called **the Markov property** of the Markov chain. Thus, by the Markov property,
  Given the state of the Markov chain at present (X_n) its future state (X_{n+1}) is independent of the past states $(X_{n-1}, ..., X_1, X_0)$.

Examples of Markov chains

We consider a stochastic process $\{X_n : n = 0, 1, 2, ...\}$ that takes on a finite or countable set M, and an element of M called state.

Example 3.1.5. Let X_n be the weather of the nth day which can be

$$M = \{sunny, windy, rainy, cloudy\}$$

Then, perhaps we have the following case:
$X_0 = sunny, X_1 = windy, X_2 = rainy, X_3 = sunny, X_4 = cloudy, ...$

Example 3.1.6. Let X_n be the product sales on the nth day which can be

$$M = \{0, 1, 2, ...\}$$

Then, perhaps we have the following case:
$X_0 = 0, X_1 = 5, X_2 = 2, X_3 = 0, X_4 = 5, ...$

Definition 3.1.7. [4] If P_{ij} represents the probability that the process will make a transition to state i given that currently the process is state j,then the matrix P_{ij}, the transition probabilities

$$\begin{pmatrix} P_{00} & P_{01} & \dots \\ P_{10} & P_{11} & \dots \\ \vdots & \vdots & \vdots \end{pmatrix}$$

is called the one-step transition probability matrix of the process.
And the entries of the transition probability matrix satisfy the following properties:

$$P_{ij} \geq 0, \qquad \sum_{i=0}^{\infty} P_{ij} = 1, \quad \forall j = 0, 1, ...$$

Example 3.1.8. (**Revisited**) In Example 3.1.1 $\{X_n : n = 0, 1, ...\}$ is a Markov chain with state space $\{1, 2, 3, 4\}$ and transition matrix

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

And satisfies

$$P_{ij} \geq 0 \qquad \forall i, j \in \{1, 2, 3, 4\}$$

and

$$\sum_{i=0}^{\infty} P_{ij} = 1 \qquad \forall j \in \{0, 1, ...\} \tag{3.5}$$

Property 3.5 is that they sum to 1 i.e.,

$$P(X_{n+1} = 1|X_n = i) + P(X_{n+1} = 2|X_n = i) + P(X_{n+1} = 3|X_n = i) + P(X_{n+1} = 4|X_n = i) = 1$$

For $i = 1$,

$$P(X_{n+1} = 1|X_n = 1) + P(X_{n+1} = 2|X_n = 1) + P(X_{n+1} = 3|X_n = 1) + P(X_{n+1} = 4|X_n = 1) = P_{11} + P_{12} + P_{13} + P_{14}$$
$$= 0 + \frac{1}{2} + 0 + \frac{1}{2} = 1.$$

Example 3.1.9. In a town there are tow supermarkets only, namely Wellcome and $Park'n$. A marketing research indicated that a consumer of Wellcome may switch to $Park'n$ in his/her next shopping with a probability of $(\alpha > 0)$, while a consumer of $Park'n$ may switch to Wellcome in his/her next shopping with a probability of $(\beta > 0)$. Let X_n be a 2-state process (taking values of $\{0, 1\}$) describing the behaviour of consumer. We have $X_n = 0$ if the consumer shops with Wellcome on the nth day and $X_n = 1$ if the consumer shops with $Park'n$ on the nth day. Since the future state (which supermarket to shop in the next time) depends on the current state only, it is a Markov chain process. It is easy to check that the transition probabilities are

$P_{00} = 1 - \alpha$, $P_{10} = \alpha$, $P_{11} = 1 - \beta$, $P_{01} = \beta$,

Then the one-step transition matrix of this process is given by

$$\begin{pmatrix} 1 - \alpha & \beta \\ \alpha & 1 - \beta \end{pmatrix}$$

Example 3.1.10. [9] In an intersection, a working traffic light will be out of order the next day with probability 0.07, and an out-of-order traffic light will be working the next day with probability 0.88. Let $X_n = 1$ if on the day n the traffic light is working; $X_n = 0$ if the traffic light is not working. then, $\{X_n : n = 0, 1, 2, ...\}$ is a Markov chain with state space $\{0, 1\}$.
Now, we can see that $P_{00} = 1 - .88 = 0.12, P_{01} = 0.88, P_{10} = 0.07, P_{11} = 1 - 0.07 = 0.93$, so, the transition probability matrix is

$$\begin{pmatrix} 0.12 & 0.88 \\ 0.07 & 0.93 \end{pmatrix}$$

Example 3.1.11. [6] **(Brand preference)**. Suppose there are three types of laundry detergent, 1, 2, and 3, and let X_n be the brand chosen on the nth purchase. Customers who try these brands are satisfied and choose the same thing again with probabilities 0.8, 0.6, and 0.4 respectively. When they change they pick one of the other two brands at random. The transition probability is

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} \end{array}$$

Example 3.1.12. (**Random Walk**)[4] Consider a person who performs a random walk on the real line with the counting numbers

Then, let $\{..., -2, -1, 0, 1, 2, ...\}$ be the state space, see Fig. 3.2. Each time the person at state i can move one step forward $(+1)$ or one step backward (-1) with probabilities $p(0 < p < 1)$ and $(1 - p)$ respectively. Therefore we

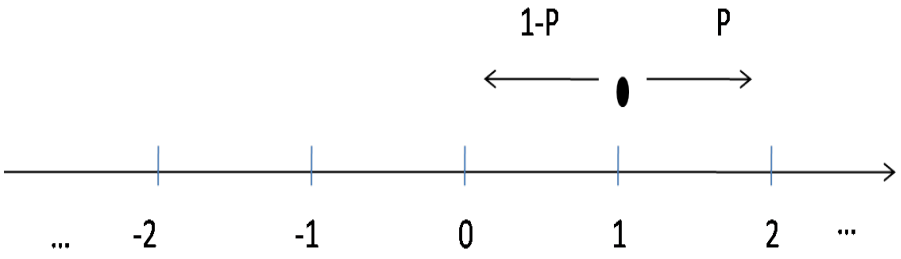

Figure 3.2: The random walk

have the transition probabilities

$$P_{ij} = \begin{cases} p & \text{if j=i+1 ,} \\ 1-p & \text{if j=i-1,} \\ 0 & \text{otherwise} \end{cases}$$

If $N = 3$, where the states are 0,1,2,3 and $p = .3$, $1 - p = .7$, then if we consider $P_{00} = P_{33} = 1$, the transition matrix will be:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ .7 & 0 & .3 & 0 \\ 0 & .7 & 0 & .3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

3.1.1 The nth Step Transition Matrix

In the previous section, we have defined the one-step transition probability matrix P for Markov chain process. In this section, we are going to investigate the n-step transition probability $P_{ij}^{(n)}$ of a Markov chain process.

Definition 3.1.13. Define $P_{ij}^{(n)}$ to be the probability that a process in state j will be in state i after n additional transitions. In particular $P_{ij}^{(1)} = P_{ij}$.

Remarks 3.1.14.

- $P^{(0)}$ is the identity matrix, that is, $p_{ij}^0 = 1$ if $i = j$, and $p_{ij}^0 = 0$ if $i \neq j$.

- Also, $P^{(1)} = P$, the transition probability matrix of Markov chain.

Theorem 3.1.15. *[9] (Chapman-Kolmogorov equation).* $\forall i, j \in S = \{x_0, x_1, ...\}$, *we have that*

$$P_{ij}^{m+n} = \sum_{k \in S} P_{ik}^{m} P_{kj}^{n}. \tag{3.6}$$

Proof. If we want to go from i to j in $m+n$ steps, we will go from i to k in m steps and from k to j in n steps, and these tow steps are independent because of Markov property.
Now,

$$\begin{aligned}
P_{ij}^{m+n} &= P(X_{m+n} = j | X_0 = i) \\
&= \sum_{k \in S} P(X_{m+n} = j, X_m = k | X_0 = i) \\
&= \sum_{k \in S} \frac{P(X_{m+n} = j, X_m = k, X_0 = i)}{P(X_0 = i)} \quad \text{(by definition of conditional probability)} \\
&= \sum_{k \in S} \frac{P(X_{m+n} = j, X_m = k, X_0 = i)}{P(X_0 = i)} \frac{P(X_m = k, X_0 = i)}{P(X_m = k, X_0 = i)} \\
&= \sum_{k \in S} \frac{P(X_{m+n} = j, X_m = k, X_0 = i)}{P(X_m = k, X_0 = i)} \frac{P(X_m = k, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k \in S} P(X_{m+n} = j | X_m = k, X_0 = i).P(X_m = k | X_0 = i) \\
&= \sum_{k \in S} P(X_{m+n} = j | X_m = k).P(X_m = k | X_0 = i) \\
&= \sum_{k \in S} P_{kj}^{n} P_{ik}^{m}
\end{aligned}$$

Note that in 3.7, P_{ij}^{n+m} is the ij th entry of the matrix $P^{(n+m)}$, P_n^{ik} is the ik th entry of the matrix $P^{(n)}$, and P_{kj}^{m} is the kj th entry of the matrix $P^{(m)}$. As we know, from the definition of the product of two matrices, the defining relation for the ij th entry of the product of matrices $P^{(n)}$ and $P^{(m)}$ is identical to 3.7. Hence the Chapman- Kolmogorov equations, in matrix

form, are

$$P^{(n+m)} = P^{(n)}P^{(m)}$$

, □

Proposition 3.1.16. *$P^{(n)} = (P)^n$, that is to say that the n-step transition probability matrix is equal to one step transition probability matrix raised to power of n.*

Proof. we will prove this proposition by using mathematical induction. when n=1, its clear that the proposition is true. Then, we assume that the proposition is true for n, i.e.,

$$P^{(n)} = (P)^n = \underbrace{P \times P \times .... \times P}_{n-times}$$

Then, we want to prove the truth of the proposition for n+1, the key in proving this is **Chapman-Kolmogorov equation**.

$$P_{ij}^{m+n} = \sum_{k \in S} P_{ik}^m P_{kj}^n \tag{3.7}$$

taking $m = n$, $n = 1$ in 3.7, we see that

$$\begin{aligned} P_{ij}^{n+1} &= \sum_{k \in S} P_{ik}^n P_{kj}^1 \\ &= [P_{ij}]^{n+1} \end{aligned}$$

□

Example 3.1.17. For the Markov chain of Example 3.1.10, the tow step transition probability matrix is given by:

$$\mathbf{P}^{(2)} = \mathbf{P}^2 = \begin{pmatrix} 0.12 & 0.88 \\ 0.07 & 0.93 \end{pmatrix} \begin{pmatrix} 0.12 & 0.88 \\ 0.07 & 0.93 \end{pmatrix} = \begin{pmatrix} 0.076 & 0.924 \\ 0.0735 & 0.9265 \end{pmatrix}$$

P_{01}^2: denote the probability that an out of order traffic light will be working the day after tomorrow, and this probability is equal to 0.924.
Simillarly, P_{10}^2 denote the probability that a working traffic light will be out of order the day after tomorrow, which has the probability 0.0735.

Remark 3.1.18. In general, $P_{ij}^n \neq (P_{ij})^n$, because that $(P_{10})^2 = (0.07)^2 = 0.0049$ not equal to $(P_{10}^2) = 0.0735$.

Theorem 3.1.19. *Let $\{X_n : n = 0, 1, ...\}$ be a Markov chain with transition probability matrix $P = (p_{ij})$. For $i \geq 0$, let $p(i) = P(X_0 = i)$ be the probability mass function of X_0. Then the probability mass function of X_n is given by*

$$P(X_n = j) = \sum_{i=0}^{\infty} P(i)P_{ij}^n, \quad j = 0, 1, 2, ...$$

Proof. Applying the law of total probability, Theorem 1.1.4, to the sequence of mutually exclusive events $\{X_0 = i\}$, $i \geq 0$, we have

$$\begin{aligned} P(X_n = j) &= \sum_{i=0}^{\infty} P(X_n = j | X_0 = i)P(X_0 = i) \\ &= \sum_{i=0}^{\infty} P_{ij}^n P(i) = \sum_{i=0}^{\infty} P(i)P_{ij}^n \end{aligned}$$

□

Example 3.1.20. [9] Suppose that a mouse is moving inside the maze shown in Figure 3.1.20, from one cell to another, in search of food. When at a cell, the mouse will move to one of the adjoining cells randomly. For $n \geq 0$, let X_n be the cell number the mouse will visit after having changed cells n times. Then $\{X_n : n = 0, 1, ...\}$ is a Markov chain with state space $\{1, 2, ..., 9\}$ and transition probability matrix

$$\begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

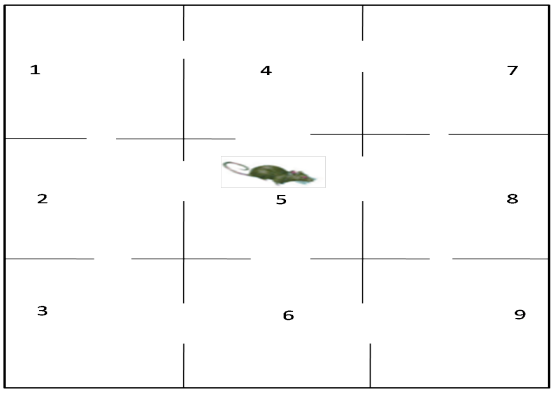

Figure 3.3: The moving mouse

Direct calculations show that the 5-step transition probability matrix for this Markov chain, is given by

$$\begin{pmatrix}
0 & \frac{5}{18} & 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 & \frac{2}{9} & 0 \\
\frac{5}{27} & 0 & \frac{5}{27} & 0 & \frac{1}{3} & 0 & \frac{4}{27} & 0 & \frac{4}{27} \\
0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 \\
\frac{5}{27} & 0 & \frac{4}{27} & 0 & \frac{1}{3} & 0 & \frac{5}{27} & 0 & \frac{4}{27} \\
0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\
\frac{4}{27} & 0 & \frac{5}{27} & 0 & \frac{1}{3} & 0 & \frac{4}{27} & 0 & \frac{5}{27} \\
0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 \\
\frac{4}{27} & 0 & \frac{4}{27} & 0 & \frac{1}{3} & 0 & \frac{5}{27} & 0 & \frac{5}{27} \\
0 & \frac{2}{9} & 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 & \frac{5}{18} & 0
\end{pmatrix}$$

This matrix shows that, for example, if the mouse is in cell 4 at a certain time, then after changing cells five times, the mouse will be in cell 5 with probability 1/3, in cell 7 with probability 5/27, and in cell 9 with probability 4/27.

Note that initially, it is equally likely that the mouse is in any of the 9 cells. That is,

$$P(i) = P(x_0 = i) = \frac{1}{9} \qquad 1 \leq i \leq 9$$

Then, using the matrix P^5 and Theorem 3.1.19, we can readily find the probability that the mouse is in cell j, $1 \leq j \leq 9$, after 5 transitions. For example,

$$\begin{aligned} P(X_5 = 4) &= \sum_{i=1}^{9} P(i) P_{i4}^5 = \frac{1}{9} \sum_{i=1}^{9} P_{i4}^5 \\ &= \frac{1}{9}(\frac{5}{18} + 0 + \frac{2}{9} + 0 + \frac{1}{4} + 0 + \frac{5}{18} + 0 + \frac{2}{9}) = 0.139 \end{aligned}$$

3.1.2 Absorbing Markov Chains

There is many special types of Markov chains. The first type that we shall study is called an absorbing Markov chain.

Definition 3.1.21. [34] A state x_i of a Markov chain is called absorbing if it is impossible to leave it (*i.e.*, $P_{ii} = 1$). A Markov chain is absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state (not necessarily in one step).

Definition 3.1.22. In an absorbing Markov chain, a state which is not absorbing is called transient.

Example 3.1.23. Drunkard's Walk

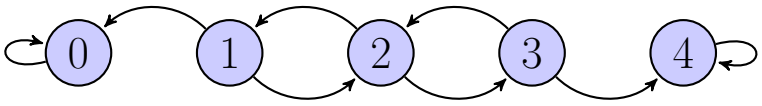

Figure 3.4: Drunkard's walk

A man walks along a four-block stretch of Park Avenue (see Figure 3.1.23). If he is at corner 1, 2, or 3, then he walks to the left or right with equal probability. He continues until he reaches corner 4, which is a bar, or corner 0, which is his home. If he reaches either home or the bar, he stays there. We form a Markov chain with states 0, 1, 2, 3, and 4. States 0 and 4 are absorbing states,since the man will stay in these states. The transition matrix is then

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The states 1, 2, and 3 are transient states, and from any of these it is possible to reach the absorbing states 0 and 4. Hence the chain is an absorbing chain. When a process reaches an absorbing state, we say that it is absorbed.

Example 3.1.24. Consider a Markov chain with transition probability matrix

$$\begin{pmatrix} 0 & \frac{2}{7} & 0 & \frac{5}{7} & 0 \\ \frac{5}{6} & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & \frac{2}{5} & \frac{3}{5} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

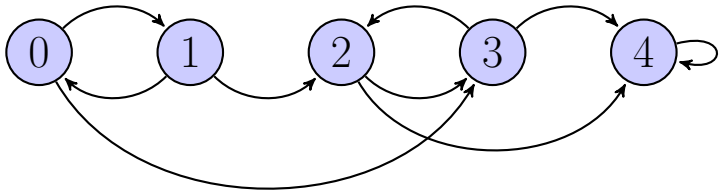

Figure 3.5: Transition graph of Example 3.1.24

Figure 3.1.24 is a transition graph for this Markov chain. It shows that states 0 and 1 communicate. So, they belong to the same class. State 2 is accessible from 1, but not vice versa. So 2 does not belong to the class of 0 and 1. States 3 and 2 communicate. Therefore, 3 does not belong to the class of 0 and 1 either. States 2 and 3 belong to the same class. State 4 is accessible from states 0, 1, 2, and 3, but no state is accessible from 4. So 4 belongs to a class by itself. Thus this Markov chain consists of three classes: $\{0, 1\}$, $\{2, 3\}$, and $\{4\}$. In this example, note that, for state 4, $P_{44} = 1$. That is, once the process enters 4, it will stay there forever. Such states are called absorbing. In general, state i of a Markov chain is absorbing if $P_{ii} = 1$.

3.1.3 Irreducibility and Aperiodicity of Markov Chains

To express the definition of irreducible Markov chains, we will introduce the meaning of communicate states.

Let $\{X_n : n = 0, 1, ...\}$ be a Markov chain with state space S and transition probability matrix P. A state j is said to be accessible from state i if there is a positive probability that, starting from i, the Markov chain will visit state j after a finite number of transitions. If j is accessible from i, we write $i \to j$. Therefore, $i \to j$ if for some $n \geq 0$, $P_{ij}^n > 0$. If i and j are accessible from each other, then we say that i and j communicate and write $i \leftrightarrow j$. Clearly, communication is a relation on the state space of the Markov chain. We will now show that this relation is an equivalence relation. That is, it is reflexive, symmetric, and transitive

Reflexivity: For all $i \in S$, $i \leftrightarrow i$ since $p_{ii}^0 = 1 > 0$.

Symmetry : If $i \leftrightarrow j$, then $j \leftrightarrow i$. This follows from the definition of i and j being accessible from each other.

Transitivity : We want to show that if $i \leftrightarrow j$, and $j \leftrightarrow k$, then $i \leftrightarrow k$. To show this, we firstly will see that $i \to k$. Now $i \to j$ implies that there exists $n \geq 0$ such that $P_{ij}^n > 0$; $j \to k$ implies that there exists $m \geq 0$ such that $p_{jk}^m > 0$. By the Chapman-Kolmogorov equations,

$$P_{ik}^{n+m} = \sum_{l=0}^{\infty} P_{il}^n P_{lk}^m \geq P_{ij}^n P_{jk}^m > 0$$

Similarly, we can see that $k \to i$. so, $i \leftrightarrow k$. As we know, an equivalence relation on a set divides that set into a collection of disjoint subsets, called equivalence classes, or simply classes. For a Markov chain, the equivalence relation defined by communication divides the state space into a collection of disjoint classes, where each class contains all of those elements of the state space that communicate with each other. Therefore, the states that communicate with each other belong to the same class. If all of the states of a Markov chain communicate with each other, then there is only one class. In such a case, the Markov chain is called **irreducible**.

Definition 3.1.25. A Markov chain is called an *irreducible* chain if it is possible to go from every state to every state (not necessarily in one move);

i.e., each state is communicate with each other. Otherwise the chain is said to be **reducible**.
Another way to express the definition would be to say that the chain is irreducible if for any states s_i , $s_j \in S$, where S is the state space, we can find an n such that $(P_{ij}^n) > 0$. An easy way to verify that a Markov chain is irreducible is to look at its transition graph, and check that from each state there is a sequence of arrows leading to any other state.

In many books, *irreducible* Markov chains are called *ergodic*.

Example 3.1.26. A reducible Markov chain. Consider a Markov chain $(X_0, X_1, ...)$ with state space $S = \{1, 2, 3, 4\}$ and transition matrix

$$\begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0.8 & 0.2 \end{pmatrix}$$

We note that if the chain starts in state 1 or state 2, then we will still in states 1 and 2 forever.
Similarly, if we start in state 3 or state 4, then we can never leave the subset $\{3, 4\}$ of the state space. Hence, the chain is reducible.

Definition 3.1.27. A Markov chain is called a *regular* chain if some power of the transition matrix has only positive elements.

In other words, for some n, it is possible to go from any state to any state in exactly n steps. It is clear from this definition that every regular chain is irreducible. On the other hand, an irreducible chain is not necessarily regular, as the following examples show.

Example 3.1.28. Let the transition matrix of a Markov chain be defined by

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

If we find the second transition matrix P^2, we note that
$P^2 = (P)^2 =$

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So, we can move from any state to any state by one step or two steps, so this chain is irreducible. However, if n is odd, then it is not possible to move from state 0 to state 0 in n steps, and if n is even, then it is not possible to move from state 0 to state 1 in n steps, so the chain is not regular.

Period[9]

Let $\{X_n : n = 0, 1, ...\}$ be a Markov chain with state space S and transition probability matrix $P = (p_{ij})$. For $i \in S$, suppose that, starting from i, there is only a positive probability to return to i after $n_1, n_2, ...$ transitions. Then $p_{ii}^{n_1} > 0$, $P_{ii}^{n_2} > 0$, ... However, $P_{ii}^{n} = 0$ if $n \notin \{n_1, n_2, ...\}$. Let d be the greatest common divisor of n_1, n_2, ... Then $d(i)$ is said to be the period of i. The period $d(i)$ of a state $i \in S$ is defined as

$$d(i) = \gcd\{n \geq 1 : P_{ii}^{n} > 0\}$$

In words, the period of i is the greatest common divisor of the set of times that the chain can return to i , given that we start with $X_0 = i$. If $d(i) = 1$, then we say that the state i is **aperiodic**.

Definition 3.1.29. [11] A Markov chain is said to be **aperiodic** if all its states are aperiodic. Otherwise the chain is said to be **periodic**.

Remark 3.1.30. The period is a class property ,that is,

if $i \leftrightarrow j$, then i and j have the same period.

Example 3.1.31. Consider a Markov chain $\{X_n : n = 0, 1, ...\}$ with state space $\{0, 1, ..., m-2, m-1\}$. Clearly, the set of all integers $n \geq 1$ for which $p_{00}^{n} > 0$ is $\{m, 2m, 3m, ...\}$. Since the greatest common divisor of this set is

m, the period of 0 is m. Since this Markov chain is irreducible, there is one class and so the period of any other state is also m.

Example 3.1.32. [18] (Random walk on the n-cycle) Let $S = \mathbb{Z}_n = \{0, 1, ..., n-1\}$, the set of remainders modulo n.

Consider the transition matrix

$$P_{jk} = \begin{cases} \frac{1}{2} & \text{if K} \equiv \text{j+1 (mod n)}, \\ \frac{1}{2} & \text{if K} \equiv \text{j-1 (mod n)}, \\ 0 & \text{otherwise} \end{cases}$$

The associated Markov chain (X_t) is called random walk on the n-cycle, this chain can be specified simply by words: at each step, a coin is tossed. If the coin lands heads up, the walk moves one step clockwise. If the coin lands tails up, the walk moves one step counterclockwise.

For $n \geq 1$, random walk on the n-cycle is irreducible, since for each $i, j \in \mathbb{Z}_n$, $i \leftrightarrow j$.

Random walk on any even length cycle is periodic, since $\gcd\{t : P^t_{x,x} > 0\} = 2$. Random walk on any odd length cycle is aperiodic.

Definition 3.1.33. For a Markov chain $\{X_n : n = 0, 1, ...\}$, let f^n_{ii} be the probability that, starting from state i, the process will return to state i, for the first time, after exactly n transitions. Let f_i be the probability that, starting from state i, the process will return to state i after a finite number of transitions. So, it is clear that $f_i = \sum_{n=1}^{\infty} f^n_{ii}$. If $f_i = 1$, then the state i is called **recurrent**. State i is called **transient** if $f_i < 1$that is, if starting from i, there is a positive probability that the process does not return to i.

Remark 3.1.34. In an irreducible Markov chain, either all states are transient, or all states are recurrent, since the transient and recurrent properties are class properties, and the irreducible Markov chain has only one class. In a

reducible Markov chain, the elements of each class are either all transient, or they are all recurrent.

Definition 3.1.35. Let i be a recurrent state of a Markov chain. The state i is called **positive recurrent** if the expected number of transitions between two consecutive returns to i is finite. If a recurrent state i is not positive recurrent, then it called **null recurrent**.

3.1.4 The Stationary Distribution of Markov chains

In this section we consider one of the most important subject in Markov theory, we will see the behavior of Markov chain after long time of transitions. Let us revisit Example 3.1.10 in which a working traffic light will be out of order the next day with probability 0.07, and an out-of-order traffic light will be working the next day with probability 0.88. Let $X_n = 1$, if on day n the traffic light will work; $X_n = 0$, if on day n it will not work. We showed that $\{X_n : n = 0, 1...\}$ is a Markov chain with state space $\{0, 1\}$ and transition probability matrix

$$P = \begin{pmatrix} 0.12 & 0.88 \\ 0.07 & 0.93 \end{pmatrix}$$

Direct calculations show that

$$P^{(6)} = P^6 = \begin{pmatrix} 0.0736842 & 0.926316 \\ 0.0736842 & 0.926316 \end{pmatrix}$$

This shows that, whether or not the traffic light is working today, the probability that it will be working after six day is 0.926316, and the probability that it will be out of order is 0.0736842. For certain Markov chains, after a large number of transitions, the probability of entering a specific state becomes independent of the initial state of the Markov chain. Mathematically,

this means that for such Markov chains $\lim_{n\to\infty} P_{ij}^n$ converges to a limiting probability that is independent of the initial state i. For some Markov chains, these limits either cannot exist or they do not converge to limiting probabilities.

In general, it can be shown that for an irreducible, positive recurrent, aperiodic Markov chain $\{X_n : n = 0, 1, ...\}$ with state space $\{0, 1, ...\}$ and transition probability matrix $P = (p_{ij})$, $\lim_{n\to\infty} p_{ij}^n$ exists and is independent of i. The limit is denoted by π_j, and $\sum_{j=0}^n \pi_j = 1$. This can expressed by symbols by:

$$\lim_{n\to\infty} P(X_n = j) = \pi_j.$$

Since, by conditioning on X_n and law of total probability, we have

$$P(X_{n+1} = j) = \sum_{i=0}^{\infty} P(X_{n+1} = j|X_n = i)P(X_n = i) = \sum_{i=0}^{\infty} p_{ij}P(X_n = i),$$

as $n \to \infty$, we must have

$$\pi_j = \sum_{i=0}^{\infty} p_{ij}\pi_i, \quad j \geq 0. \tag{3.8}$$

This system of equations along with $\sum_{j=0}^{\infty} \pi_j = 1$ enable us to find the limiting probabilities π_j. Let

$$\Pi = \begin{pmatrix} \pi_0 \\ \pi_1 \\ \vdots \end{pmatrix}$$

, and let P^T be the transpose of the transition probability matrix P; then equation 3.1.4 in matrix form are

$$\Pi = P^T\Pi$$

In particular, for a finite-state Markov chain with space $\{0, 1, ...\}$, this equation is

$$\begin{pmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_n \end{pmatrix} = \begin{pmatrix} p_{00} & p_{10} & \dots & p_{n1} \\ p_{01} & p_{11} & \dots & p_{n1} \\ \vdots & & & \\ p_{0n} & p_{1n} & \dots & p_{nn} \end{pmatrix} \begin{pmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_n \end{pmatrix}$$

If for a Markov chain, for each $j \geq 0$, $\lim_{n\to\infty} p_{ij}^n$ exists and is independent of i, we say that the Markov chain is in **equilibrium** or **steady state**. The limits $pi_j = \lim_{n\to\infty} p_{ij}^n, j \geq 0$, are called the **stationary probabilities**.

Definition 3.1.36. [11] Let $\{X_0, X_1, ...\}$ be a Markov chain with state space $\{s_1, ..., s_k\}$ and transition matrix P. A row vector $\pi = (\pi_1, ..., \pi_k)$ is said to be a stationary distribution for the Markov chain, if it satisfies

- $\pi_i \geq 0$ for $i = 1, ..., k$, and $\sum_{i=1}^k \pi_i = 1$, and
- $\pi P = \pi$, meaning that $\sum_{i=1}^k \pi_i P_{ij} = \pi_j$ for $j = 1, ..., k$.

Theorem 3.1.37. *[9] Let $\{X_n : n = 0, 1, ...\}$ be an irreducible, positive recurrent, aperiodic Markov chain with state space $\{0, 1, ...\}$ and transition probability matrix $P = (p_{ij})$. Then, for each $j \geq 0$, $\lim_{n\to\infty} p_{ij}^n$ exists and is independent of i. Let $\pi_j = \lim_{n\to\infty} p_{ij}^n$, $j \geq 0$ and $\pi = (\pi_0 \pi_1 \ldots)^T$ .We have*

(a) $\Pi = P^T \Pi$, *and* $\sum_{j=0}^{\infty} \pi_j = 1$. *Furthermore, these equations determine the stationary probabilities, $\pi_0, \pi_1, ...,$ uniquely.*

(b) π_j *is the long-run proportion of the number of transitions to state j , $j \geq 0$.*

(c) *The expected number of transitions between two consecutive visits to state j is $1/\pi_j$, $j \geq 0$.*

Example 3.1.38. An engineer analyzing a series of digital signals generated by a testing system observes that only 1 out of 15 highly distorted signals follows a highly distorted signal, with no recognizable signal between, whereas 20 out of 23 recognizable signals follow recognizable signals, with no highly distorted signal between. Given that only highly distorted signals are not recognizable, find the fraction of signals that are highly distorted.

solution: For $n \geq 1$, let $X_n = 1$, if the nth signal generated is highly distorted; $X_n = 0$, if the nth signal generated is recognizable. Then $\{X_n : n = 0, 1, ...\}$ is a Markov chain with state space $S = \{0, 1\}$ and transition probability matrix

$$P = \begin{pmatrix} 20/23 & 3/23 \\ 14/15 & 1/15 \end{pmatrix}$$

Now, note that, the Marcov chain is irreducible, since $P_{ij} > 0 \ \forall i, j \in S$, positive recurrent because the expected number of transitions between two consecutive returns to i is finite, and aperiodic. Let π_0 be the fraction of signals that are recognizable, and π_1 be the fraction of signals that are highly distorted. Then by Theorem 3.1.37, π_0 and π_1 satisfy

$$\begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix} = \begin{pmatrix} 20/23 & 14/15 \\ 3/23 & 1/15 \end{pmatrix} \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix}$$

which gives the following system of equations:

$$\begin{cases} \pi_0 = \frac{20}{23}\pi_0 + \frac{14}{15}\pi_1 \\ \pi_1 = \frac{3}{23}\pi_0 + \frac{1}{15}\pi_1 \end{cases}$$

The first equation of the previous equations with $\pi_0 + \pi_1 = 1$, we have that

$$1 - \pi_1 = \frac{20}{23}(1 - \pi_1) + \frac{14}{15}\pi_1$$

$$\text{then,} \quad \pi_1 - \frac{20}{23}\pi_1 + \frac{14}{15}\pi_1 = \frac{3}{23}$$

Then, $\pi_1 \approx 0.123$, it yields that $\pi_0 \approx 0.877$. Therefor, approximately 12.3% of the signals generated by the testing system are highly distorted.

Definition 3.1.39. A stationary Markov chain with a transition matrix P and stationary distribution π is called **reversible**, if for arbitrary $n \geq 0$ and $x_0, x_1, ..., x_n \in S$,

$$P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = P(X_n = x_0, X_{n-1} = x_1, ..., X_0 = x_n) \tag{3.9}$$

Definition 3.1.40. [11] Suppose a probability π on S satisfies

$$\pi_i P_{ij} = \pi_j P_{ji} \qquad \forall i, j \in S. \tag{3.10}$$

The equation 3.10 the **detailed balance equation**. And π is said to be reversible if it satisfies the detailed balance equation.

Example 3.1.41. [6] (**Revisited**) In Example 3.1.11 X_n is a Markov chain with a probability transition matrix

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

Let $\Pi = (\pi_1, \pi_2, \pi_3)$ be the fraction of the three types of laundry detergent $1, 2, 3$ respectively.
Then, equation $\pi P = \pi$ yields that,

$$(\pi_1, \pi_2, \pi_3) \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = (\pi_1, \pi_2, \pi_3)$$

which gives the following system of equations:

$$\begin{cases} 0.8\pi_1 + 0.2\pi_2 + 0.3\pi_3 = \pi_1 \\ 0.1\pi_1 + 0.6\pi_2 + 0.3\pi_3 = \pi_2 \\ 0.1\pi_1 + 0.2\pi_2 + 0.4\pi_3 = \pi_3 \end{cases}$$

if we solve the previous equations with $\pi_1 + \pi_2 + \pi_3 = 1$, we get that:

$$\pi_1 = \frac{6}{11}, \quad \pi_2 = \frac{3}{11}, \quad \pi_3 = \frac{2}{11}.$$

Therefore the stationary distribution is given by:

$$\pi = \left(\frac{6}{11}, \frac{3}{11}, \frac{2}{11}\right).$$

since,

$$\pi_1 P_{12} = \pi_2 P_{21} = \frac{6}{110}$$
$$\pi_1 P_{13} = \pi_3 P_{31} = \frac{6}{110}$$
$$\pi_2 P_{23} = \pi_3 P_{32} = \frac{6}{110}$$

then the detailed balance equation (reversibility) holds.

Theorem 3.1.42. *[18] Let P be the transition matrix of a Markov chain with state space. Any distribution π satisfying the detailed balance equations 3.10 is stationary for P.*

For the proof of this theorem and more details, you can see

But the converse of this theorem is not necessary true, and the following example show this:

Example 3.1.43. Let X_n be a familys social class in the *nth* generation, which we assume is either 1 = lower, 2 = middle, or 3 = upper. In our simple version of sociology, changes of status are a Markov chain with the following transition matrix

$$P = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

Let $\pi = (pi_1, \pi_2, \pi_3)$ be the fraction of the lower, middle, and upper social class respectively. Then the equation $\pi P = \pi$, gives that:

$$(\pi_1, \pi_2, \pi_3) \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} = (\pi_1, \pi_2, \pi_3)$$

So, we get the following system of equations:

$$\begin{cases} 0.7\pi_1 + 0.3\pi_2 + 0.2\pi_3 = \pi_1 \\ 0.2\pi_1 + 0.5\pi_2 + 0.4\pi_3 = \pi_2 \\ 0.1\pi_1 + 0.2\pi_2 + 0.4\pi_3 = \pi_3 \end{cases}$$

Then, by solving these equations with $\pi_1 + \pi_2 + \pi_3 = 1$, we get that, the stationary distribution is given by:

$$\pi = \left(\frac{22}{47}, \frac{16}{47}, \frac{9}{47}\right)$$

To check this we note that

$$\begin{pmatrix} \frac{22}{47} & \frac{16}{47} & \frac{9}{47} \end{pmatrix} \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} = \begin{pmatrix} \frac{22}{47} & \frac{16}{47} & \frac{9}{47} \end{pmatrix}$$

since

$$\begin{aligned} \pi_1 P_{12} &= \frac{22}{47} . \frac{2}{10} \\ &= \frac{44}{470} \end{aligned}$$

and,

$$\begin{aligned} \pi_2 P_{21} &= \frac{16}{47} . \frac{3}{10} \\ &= \frac{48}{470} \end{aligned}$$

then,

$$\pi_1 P_{12} \neq \pi_2 P_{21}$$

Therefore, the reversibility fails.

3.2 Markov Chain Monto Carlo Methods (MCMC)

In this section, we will look at Markov chain Monte Carlo (MCMC) methods for generating samples from the posterior distribution. we construct a Markov chain that has the posterior distribution as its stationary distribution. The Metropolis-Hastings (M-H) algorithm, Gibbs sampler are methods of doing this.

3.2.1 Introduction to MCMC

In Bayesian statistics, we have two sources of information about the parameter θ: our prior belief and the observed data. The prior distribution summarizes our belief about the parameter before we look at the data. The prior density $g(\theta)$ gives the relative belief weights we have for all possible values of the parameter θ before we look at the data, and all the information about the parameter θ that is in the observed data y is contained in the likelihood function $f(y|\theta)$. However, the parameter is considered a random variable, so the likelihood function is written as a conditional distribution. The likelihood function gives the relative support weight each possible value of the parameter θ has from the observed data.
Bayes' Theorem combines the two sources into a single relative belief weight distribution after we have observed the data. The final belief weight distribution is known as the posterior distribution and it takes into account both the prior belief and the support from the data. Bayes' Theorem express the form of posterior to be proportional to prior times likelihood. In equation form this is

$$g(\theta|y) \propto g(\theta) \times f(y|\theta) \tag{3.11}$$

This formula does not give the posterior density $g(\theta|y)$ exactly, but it does give its shape, but the actual posterior density is found by scaling it so it

integrates to one.

$$g(\theta|y) = \frac{g(\theta) \times f(y|\theta)}{\int g(\theta) \times f(y|\theta) d\theta} \tag{3.12}$$

To find the actual posterior, this requires integrating

$$\int f(y|\theta) g(\theta) d\theta$$

numerically, which may be very difficult. To do this, we draw a Monte Carlo sample from the posterior. A Monte Carlo random sample from the posterior will approximate the true posterior when the sample size is large enough.
The idea of the MCMC is the following: Suppose we can construct an irreducible and aperiodic Markov chain $(X_0, X_1, ...)$, whose (unique) stationary distribution is π. If we run the chain with arbitrary initial distribution (for instance, starting in a fixed state), then the Markov chain convergence theorem (Theorem 3.1.37) guarantees that the distribution of the chain at time n converges to π, as $n \to \infty$. Hence, if we run the chain for a sufficiently long time n, then the distribution of X_n will be very close to π.
Thus, we will set the long-run distribution $\pi(\theta)$ for the Markov chain equal to the posterior density $g(\theta|y)$. Generally we will only know the not exact posterior density $g(\theta|y) \propto g(\theta) \times f(y|\theta)$. Fortunately, we will see that the not exact posterior is all we need to know to find a Markov chain that has the exact posterior as its long-run distribution.[2]

3.2.2 The Metropolis-Hastings Algorithm

This section will introduce one of the MCMC methods: the Metropolis-Hastings algorithm, which goes back to Metropolis et al. (1953) and Hastings (1970). The Metropolis-Hastings algorithm is based on proposing values sampled from an instrumental distribution, which are then accepted with a certain probability that reflects how likely it is that they are from the target distribution f.

The Metropolis-Hastings algorithm is the most popular example of the MCMC methods. Suppose we have posterior distribution π, if we want to sample from π, then we construct a Markov chain whose stationary distribution is π, and run a Markov chain long enough and then use Metropolis-Hastings algorithm.
Now, we show how the Metropolis-Hastings algorithm can be used to find a Markov chain that has the posterior as its long-run distribution.
The algorithm proceeds as follows:

1. Select the proposal distribution $q(x, y)$ that is easy to sample from.
2. Select starting point $x = x_0 \sim q(x, y)$.
3. Generate candidate point $x^* \sim q(x, x^*)$ and $u \sim uniform(0, 1)$.
4. Calculate the acceptance probability α, which is given by:

$$\alpha = \min\left\{1, \frac{\pi(x^*)q(x^*, x)}{\pi(x)q(x, x^*)}\right\}$$

5. We now either accept x^* or reject it as follows

$$X_{n+1} = \begin{cases} x^* & \text{, if } u \leq \alpha \\ x_n & \text{, otherwise.} \end{cases}$$

Repeat steps (3), (4), and (5), this generates a sequence of sample.

If the proposal distribution is symmetric,

$$q(x, y) = q(y, x)$$

we obtain the Metropolis algorithm. In this case the acceptance probability

$$\alpha = \min\left\{1, \frac{\pi(x^*)}{\pi(x)}\right\}$$

Since a move is accepted with probability

$$\alpha(x,y) = \min\left\{1, \frac{\pi(x^*)q(x^*,x)}{\pi(x)q(x,x^*)}\right\}$$

so the transition probability

$$p(x,y) = q(x,y)\alpha(x,y)$$

To check that π satisfies the detailed balance condition, we will introduce this theorem

Theorem 3.2.1. *The Metropolis algorithm produce a Markov chain* $\{X_0, X_1, X_2, ...\}$, *which is reversible with respect to stationary distribution* $\pi(x)$.

Proof. Let the proposal distribution is $q(x,y)$, and the acceptance probability is α.

Set $P(x,y) = q(x,y)\alpha$ to construct the transition probabilities.
We must show that $\pi(x)P(x,y) = \pi(y)P(y,x)$.
Obviously this holds if $x = y$. We will consider $x \neq y$, then

$$\begin{aligned}
\pi(x)P(x,y) &= \pi(x)q(x,y)\alpha \\
&= \pi(x)q(x,y)\min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \\
&= \min\left\{\pi(x)q(x,y), \pi(x)q(x,y)\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \\
&= \min\left\{\pi(x)q(x,y), \pi(y)q(y,x)\right\} \qquad (*)
\end{aligned}$$

and,

$$\begin{aligned}
\pi(y)P(y,x) &= \pi(y)q(y,x)\alpha \\
&= \pi(y)q(y,x)\min\left\{1, \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\right\} \\
&= \min\left\{\pi(y)q(y,x), \pi(y)q(y,x)\frac{\pi(y)q(x,y)}{\pi(y)q(y,x)}\right\} \\
&= \min\left\{\pi(y)q(y,x), \pi(x)q(x,y)\right\} \qquad (**)
\end{aligned}$$

from $(*)$ and $(**)$ we obtain that

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

Therefore, the chain is reversible and satisfies the detailed balance condition with respect to stationary distribution $\pi(x)$. □

This example explains the Metropolis algorithm manually:

Example 3.2.2. Write the Metropolis algorithm for obtaining samples from the posterior distribution

$$\pi(\theta) = g(\theta|y) = 0.6e^{-\frac{1}{2}\theta^2} + 0.4 \times \frac{1}{2}e^{-\frac{1}{2\times 2^2}(\theta-3)^2}$$

Which is a mixture of normal $(0, 1^2)$ and normal $(3, 2^2)$, and this is only the unscaled target since multiplying by a constant would multiply both the numerator and denominator by the constant which would cancel out.

We will use the candidate (proposal) density with variance $\sigma^2 = 1$ centered around the current value. Its shape is given by

$$q(\theta, \theta') = e^{-\frac{1}{2}(\theta'-\theta)^2}$$

Let the starting value be $\theta = 2$, since the candidate density is symmetric about the current value, then $q(\theta, \theta') = q(\theta', \theta)$, and the acceptance probability

$$\alpha = \min\left\{1, \frac{g(\theta'|y)q(\theta', \theta)}{g(\theta|y)q(\theta, \theta')}\right\}$$

So,

$$\alpha = \min\left\{1, \frac{g(\theta'|y)}{g(\theta|y)}\right\}$$

The Metropolis-Hastings algorithm proceeds as follow:
Let starting value $\theta_0 = 2$.

Draw $\theta' = 1.55$ from $q(\theta^{(n-1)}, \theta') = q(\theta^{(0)}, \theta') = q(2, \theta') = e^{-\frac{1}{2}(\theta'-2)^2}$.
Calculate the probability $\alpha(2, \theta')$.

$$\begin{aligned}
\alpha(2, 1.55) &= \min\left[1, \frac{g(\theta'|x)}{g(\theta|x)}\right] \\
&= \min\left[1, \frac{g(1.55|x)}{g(2|x)}\right] \\
&= \min\left[1, \frac{0.6 \times e^{-\frac{1}{2}\times 1.55^2} + 0.4 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(1.55-3)^2}}{0.6 \times e^{-\frac{1}{2}\times 2^2} + 0.4 \times \frac{1}{2} \times e^{-\frac{1}{2\times 2^2}(2-3)^2}}\right] \\
&= \min[1, 1.298] \\
&= 1
\end{aligned}$$

Draw $u = 0.354$ from $U(0, 1)$.
$u < \alpha(2, 1.55)$, so let $\theta^{(1)} = 1.55$.
Now, start with $\theta^{(1)} = 1.55$.
Draw $\theta' = 2.692$ from $q(\theta^{(n-1)}, \theta') = q(\theta^{(1)}, \theta') = q(1.55, \theta') = e^{-\frac{1}{2}(\theta'-1.55)^2}$.
Calculate $\alpha(1.55, 2.692) = \min[1, \frac{g(1.55|x)}{g(2.692|x)}]$

$$\begin{aligned}
\alpha(1.55, 2.692) &= \min\left[1, \frac{0.6 \times e^{-\frac{1}{2}\times 1.55^2} + 0.4 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(1.55-3)^2}}{0.6 \times e^{-\frac{1}{2}\times 2.692^2} + 0.4 \times \frac{1}{2} \times e^{-\frac{1}{2\times 2^2}(2.692-3)^2}}\right] \\
&= min[1, 1.5651] \\
&= 1
\end{aligned}$$

We can continue this process, and record the values of current value of θ.

3.2.3 The Gibbs Sampler

The Gibbs sampler is one way of MCMC methods, which help us to generate samples from joint (posterior) distributions. In this method, the samples do not generate directly from the joint (posterior) distribution, but generate from the conditional distributions derived from the joint (posterior) distribution.

Gibbs sampler is a special case of the Metropolis Hasting, in such case the acceptance probability α will be 1, so the candidate will be accepted at each step since the candidates are being drawn from the correct full conditional distribution. [2].

To introduce the Gibbs sampler, let π be a joint (posterior) distribution of a bivariate random vector (X, Y). Let $\pi(X|Y)$ be the conditional probability distribution of X given Y. Similarly, let $\pi(Y|X)$ be the conditional probability distribution of Y given X.

Now generate a bivariate Markov chain $Z_n = (X_n, Y_n)$ as follows:

Start with some $X_0 = x_0$,

$$X_k \sim \pi(X|Y_{k-1}), \qquad \text{for k= 1, 2, ...} \tag{3.13}$$

$$Y_k \sim \pi(Y|X_k), \qquad \text{for k= 0, 1, 2, ...} \tag{3.14}$$

The next example explains the Gibbs sampler manually:

Example 3.2.3. Suppose the joint (posterior) distribution of $x = 0, 1, ..., n$ and $0y1$ is given by:

$$\pi(x, y) = \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \tag{3.15}$$

Now we need to calculate the marginal distribution of x and the marginal distribution of y as follows: The marginal distribution of x is given by:

$$\begin{aligned}
\pi(x) &= \int_0^1 \pi(x, y)dy \\
&= \int_0^1 \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}dy \\
&= \frac{n!}{(n-x)!x!} \int_0^1 y^{x+\alpha-1}(1-y)^{n-x+\beta-1}dy \\
&= \frac{n!}{(n-x)!x!} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}
\end{aligned}$$

The marginal distribution of y is given by:

$$\begin{aligned}
\pi(y) &= \sum_{x=0}^{n} \pi(x,y) \\
&= \sum_{x=0}^{n} \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \\
&= y^{\alpha-1}(1-y)^{\beta-1} \sum_{x=0}^{n} \frac{n!}{(n-x)!x!} y^{x}(1-y)^{n-x} \\
&= y^{\alpha-1}(1-y)^{\beta-1}.(1). \\
&= y^{\alpha-1}(1-y)^{\beta-1}.
\end{aligned}$$

The conditional probability distribution of x given y is given by:

$$\begin{aligned}
\pi(x|y) &= \frac{\pi(x,y)}{\pi(y)} \\
&= \frac{\frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}}{y^{\alpha-1}(1-y)^{\beta-1}} \\
&= \frac{n!}{(n-x)!x!} y^{x+\alpha-1-\alpha+1}(1-y)^{n-x+\beta-1-\beta+1} \\
&= \frac{n!}{(n-x)!x!} y^{x}(1-y)^{n-x}.
\end{aligned}$$

Thus,

$$x|y \sim Bi(n,y).$$

The conditional probability distribution of y given x is given by:

$$\begin{aligned}
\pi(y|x) &= \frac{\pi(x,y)}{\pi(x)} \\
&= \frac{\frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}}{\frac{n!}{(n-x)!x!} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}} \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}.
\end{aligned}$$

Thus,

$$y|x \sim Be(x+\alpha, n-x+\beta).$$

Now generate a bivariate Markov chain $z_n = (x_n, y_n)$ as follows:
Start with some $X_0 = x_0$,

$$x_k \sim Bi(n, y_{k-1}), \qquad \text{for k} = 1, 2, ... \tag{3.16}$$

$$y_k \sim Be(x_k+\alpha, n-x_k+\beta), \qquad \text{for k} = 0, 1, 2, ... \tag{3.17}$$

To illustrate the Gibbs sampler for the above, suppose $n = 10$, $\alpha = 1$ and $\beta = 2$. The algorithm of the sampler is as follows:

- Start with $x_0 = 2$ and use it to obtain y_0 from (3.17):

$$\begin{aligned} y_0 &\sim Be(x_0+1, 10-x_0+2) \\ &= Be(3, 10), \end{aligned}$$

  which gives $y_0 = 0.2379$.
  Therefore $(x_0, y_0) = (2, 0.2379)$.

- x_1 is obtained from (3.16):

$$\begin{aligned} x_1 &\sim Bi(10, y_0) \\ &= Bi(10, 0.2379) \end{aligned}$$

  which gives $x_1 = 2$.
  y_1 is obtained from (3.17):

$$\begin{aligned} y_1 &\sim Be(x_1+1, 10-x_1+2) \\ &= Be(3, 10) \end{aligned}$$

  which gives $y_1 = 0.1334$.
  Therefore $(x_1, y_1) = (2, 0.1334)$.

- x_2 is obtained from (3.16):

$$\begin{aligned} x_2 &\sim Bi(10, y_1) \\ &= Bi(10, 0.1334), \end{aligned}$$

giving that $x_2 = 3$.
And y_2 is obtained from (3.17)

$$\begin{aligned} y_2 &\sim Be(x_2 + 1, 10 - x_2 + 2) \\ &= Be(4, 9) \end{aligned}$$

giving $y_2 = 0.4735$.
Therefore $(x_2, y_2) = (3, 0.4735)$

- x_3 is obtained from (3.16)

$$\begin{aligned} x_3 &\sim Bi(10, y_2) \\ &= Bi(10, 0.4735) \end{aligned}$$

giving $x_3 = 6$
And y_3 is obtained from (3.17)

$$\begin{aligned} y_3 &\sim Be(x_3 + 1, 10 - x_3 + 2) \\ &= Be(7, 6) \end{aligned}$$

giving $y_3 = 0.6338$
Therefore $(x_3, y_3) = (6, 0.6338)$.

- x_4 is obtained from (3.16)

$$\begin{aligned} x_4 &\sim Bi(10, y_3) \\ &= (10, 0.6338), \end{aligned}$$

giving $x_4 = 4$.

And y_4 is obtained from (3.17):

$$\begin{aligned} y_4 &\sim Be(x_4 + 1, 10 - x_4 + 2) \\ &= Be(5, 8) \end{aligned}$$

giving $y_4 = 0.4196$.

Therefore $(x_4, y_4) = (4, 0.4196)$.

The Gibbs sequence within five terms:
$(2, 0.2379), (2, 0.1334), (3, 0.4735), (6, 0.6338), (4, 0.4196), ...$

Chapter 4

Bayesian Analysis of Finite Exponential Mixtures

Exponential is useful and widely used distribution, it plays an important role in modeling continuous data, and in the analysis of failure data when the population is homogeneous. However, the world is producing more and more data with complex structure, since in many practical problems, the real data can be seen as coming from several subpopulations and the homogeneity assumption may be unsuitable in those data. Then, when the population consists of subpopulations, we prefer to use exponential mixture instead of homogeneous exponential for the data.

In this chapter we present the finite Exponential mixture model using the missing data formulation, and we will derive the full conditional posterior distributions of all parameters.

We will use Gibbs sampler and algorithm to draw samples from posterior of the exponential mixture in order to use them in the bayesian analysis.

We use these samples in the estimation of the unknown parameters of the of the model.

4.1 Finite Exponential Mixture Model

Recall that a random variable X has a finite mixture distribution if its pdf function f satisfies

$$f(x) = \sum_{j=1}^{k} p_j f_j(x)$$

Where f_j in this case are the pdfs of exponential distribution with distinct parameters of the k mixture components or populations, and the mixture proportions p_j satisfy $p_j > 0$ for $j = 1, 2, ..., k$, and $\sum_{j=1}^{k} p_j = 1$.
The probability function of the k-finite exponential mixture is given by

$$f(x|\lambda, p) = \sum_{j=1}^{k} p_j \lambda_j e^{-\lambda_j x} \tag{4.1}$$

Where $p = (p_1, p_2, ..., p_k)$, for some probabilities $p_j > 0$, $j = 1, ..., k$, $k > 1$, with $\sum_{j=1}^{k} p_j = 1$. $\lambda = (\lambda_1, \lambda_2, ..., \lambda_k)$, and we assume that $\lambda_1 < \lambda_2 < ... < \lambda_k$ to insure the identifiability of the above finite mixture.
Now, we will introduce definitions of some distributions that we will use in this section.

Definition 4.1.1. (**Bernoulli distribution**) the Bernoulli distribution, named after Swiss scientist Jacob Bernoulli, is the probability distribution of a random variable which takes the value 1 with success probability of p, and the value 0 with failure probability of $q = 1 - p$. It can be used to represent a coin toss where 1 and 0 would represent "head" and "tail" (or vice versa), respectively. The Bernoulli distribution is a special case of the binomial distribution; the Bernoulli distribution is a binomial distribution where $n = 1$.
If X is a random variable with this distribution, we have:

$$P(X = 1) = 1 - P(X = 0) = 1 - q = p$$

The probability mass function of this distribution, over possible outcomes k,

is

$$f(k|p) = \begin{cases} p & \text{if } k = 1 \ , \\ 1-p & \text{if } k = 0. \end{cases}$$

Definition 4.1.2. [40] (**Multinomial distribution**) the multinomial distribution is a generalization of the binomial distribution. For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

The binomial distribution is the probability distribution of the number of successes for one of just two categories in n independent Bernoulli trials, with the same probability of success on each trial. In a multinomial distribution, the analog of the Bernoulli distribution is the categorical distribution, where each trial results in exactly one of some fixed finite number k possible outcomes, with probabilities $p_1, ..., p_k$ (so that, $p_i \geq 0$, for $i = 1, ..., k$, and $\sum_{i=1}^{k} p_i = 1$), and there are n independent trials. Then if the random variables X_i indicate the number of times outcome number i is observed over the n trials, the vector $X = (X_1, ..., X_k)$ follows a multinomial distribution with parameters n and p, where $p = (p_1, ..., p_k)$. While the trials are independent, their outcomes X are dependent because they must be summed to n.

The probability mass function of this multinomial distribution is:

$$f(x_1, ..., x_k|n, p_1, ..., p_k) = \begin{cases} \frac{n!}{x_1!...x_k!} p_1^{x_1}...p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \ , \\ 0 & \text{otherwise.} \end{cases}$$

The probability mass function can be expressed using the gamma function as:

$$f(x_1, ..., x_k, p_1, ..., p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}$$

Definition 4.1.3. [39] (**Dirichlet Distribution**) Is a family of continuous multivariate probability distributions parameterized by a vector of positive reals. It is the multivariate generalization of the beta distribution. Dirichlet distributions are very often used as prior distributions in Bayesian statistics, and in fact the Dirichlet distribution is the conjugate prior of the multinomial distribution.

The Dirichlet distribution of order $k \geq 2$ with parameters $\delta_1, ..., \delta_k > 0$ has a probability density function

$$f(p_1, ..., p_k | \delta_1, ..., \delta_k) = \frac{1}{B(\delta)} \prod_{j=1}^{k} p_j^{\delta_j - 1}$$

where $p_j \in (0, 1)$ and $\sum_{j=1}^{k} p_j = 1$, $k \geq 2$, where $\delta_j > 0$.

The normalizing constant $B(\delta)$ is the multinomial Beta function, which can be expressed in terms of gamma function:

$$B(\delta) = \frac{\prod_{i=1}^{k} \Gamma(\delta_i)}{\Gamma(\sum_{i=1}^{k} \delta_i)}, \qquad \delta = (\delta_1, ..., \delta_k).$$

4.1.1 The Likelihood Density

Throughout our discussion, n will denote the number of data points and k will denote the number of components in the mixture formulation.

Since we present the finite exponential mixture model by using the missing data formulation, so firstly, we will introduce the missing data indicators z_i, $i = 1, 2, ..., n$.

For each observation x_i, $i = 1, ..., n$ we have an indicator z_i such that

$$z_i = (z_{ij})_{j=1}^{k} = (z_{i1}, z_{i2}, ..., z_{ik})$$

where

$$z_{ij} = \begin{cases} 1 & \text{if the observation } x_i \text{ belongs to the } j^{th} \text{ component of the mixture,} \\ 0 & \text{otherwise.} \end{cases}$$

each z_{ij} takes on two values only 1 or 0, and for each z_i only one of $z_{ij}^{,}s$ equal to 1, and the rest are all 0, therefore for fixed i, $\sum_{j=1}^{k} z_{ij} = 1$.
For each z_i we have a single trial results in exactly one of k possible components of the mixture, with probabilities $p_1, ..., p_k$, $p_j \in (0,1)$ for $j = 1, 2, ..., k$ and $\sum_{j=1}^{k} p_j = 1$.
Thus the density $f(x_i|z_{ij} = 1)$ is exponential(λ_j), and f($z_{ij} = 1|p$)=p_j.
Also for fixed i, and for all $j = 1, ..., k$, since z_{ij} takes in two values only 1 or 0, then $f(z_{ij} = 0|p) = 1 - f(z_{ij} = 1|p) = 1 - p_j$. Therefore, $z_{ij} \sim Bernolli(p_j)$, for each $z_i = (z_{ij})_{j=1}^{k}$, we have $z_i|p \sim multinomial(1, p_1, ..., p_k)$.

So, the density of the indicator $z_i = (z_{ij})_{j=1}^{k}$ is

$$\begin{aligned} f(z_{i1}, ..., z_{ik}|p_1, ..., p_k) &= \frac{1!}{z_{i1}!...z_{ik}!} \prod_{j=1}^{k} p_j^{z_{ij}} \\ &= \prod_{j=1}^{k} p_j^{z_{ij}} \qquad \text{(since the value of } z_{ij} \text{ takes only 0 or 1 for all } j = 1, ..., k) \end{aligned}$$

since $z_i^{,}s$ are independent, and by definition 1.2.2, the joint indicator density is:

$$f(z|p) = \prod_{i=1}^{n} f(z_i|p) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_j^{z_{ij}}.$$

Let $X = X_1, X_2, ..., X_n$ be an iid random sample from an exponential mixture density.

The likelihood density of the mixture is:

$$\begin{aligned} f(x|\lambda, p) &= \prod_{i=1}^{n} f(x_i|\lambda, p) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{k} p_j \frac{1}{\lambda_j} e^{-\frac{x}{\lambda_j}} \end{aligned}$$

And by using the indicator z_i we can rewrite the likelihood as

$$\begin{aligned} f(x,z|\lambda,p) &= f(x|z,\lambda,p)f(z|p,\lambda,x) \\ &= f(x|z,\lambda)f(z|p) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{1}{\lambda_j}e^{-\frac{x_i}{\lambda_j}}\right)^{z_{ij}}\prod_{i=1}^{n}\prod_{j=1}^{k}(p_j)^{z_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k}\left(p_j\frac{1}{\lambda_j}e^{-\frac{x_i}{\lambda_j}}\right)^{z_{ij}}. \end{aligned}$$

4.1.2 Priors Densities

Priors densities of the parameters are chosen to be conjugate priors.

For the weights p

We follow the classical choice of a Dirichlet prior with a parameter $\delta = (\delta_1, ..., \delta_k)$, and we assume that $\delta_j = 1$ for all $j = 1, ..., k$ (as chosen by Viallefont, V. and others.).

Let $\pi(p)$ be the prior density of proportions p of our exponential mixture. Then, we will assume that, $p \sim Dir(p_1, ..., p_k, \delta_1, ..., \delta_k)$, with $\delta_j = 1$, $\forall j = 1, ..., k$.

So the density of p is

$$\begin{aligned} \pi(p_1, ..., p_k|\delta_1, ..., \delta_k) &= \frac{1}{B(\delta)}\prod_{j=1}^{k}p_j^{\delta_j-1} \\ &= \frac{\Gamma(\sum_{j=1}^{k}1)}{\prod_{j=1}^{k}\Gamma(1)}\prod_{j=1}^{k}p_j^0 \\ &= \frac{\Gamma(k)}{\prod_{j=1}^{k}\Gamma(1)} \\ &= \frac{(k-1)!}{\prod_{j=1}^{k}1} = (k-1)! \end{aligned}$$

Note that we get the fourth equality by using the identity $\Gamma(n) = (n-1)!$, when n is a positive integer.

Now we want to prove that the conjugate prior of a mulinomial parameter $p = (p_1, ..., p_k)$ is $Dirichlet(\delta)$

If $X_1, X_2, ..., X_n$ are iid multinomial$(1, p_1, ..., p_k)$, then the density of each X_i, $i = 1, ..., n$ is

$$f(x_{i1}, ..., x_{ik}|p_1, ..., p_k) = \frac{\Gamma(\sum_{j=1}^{k} x_{ij} + 1)}{\prod_{j=1}^{k} \Gamma(x_{ij} + 1} \prod_{j=1}^{k} p_j^{x_{ij}}$$

where $x_{ij} \in \{0, 1\}$, and $\sum_{j=1}^{k} x_{ij} = 1$.

Note that, we have a single trial resulting in exactly one of some fixed finite number k possible outcomes, with probabilities $p_1, ..., p_k$.

And suppose the prior distributed as $Dirichlet(\delta)$, that is, the prior density is given by

$$f(p_1, ..., p_k|\delta_1, ..., \delta_k) = \frac{\Gamma(\sum_{j=1}^{k} \delta_j)}{\prod_{j=1}^{k} \Gamma(\delta_j)} \prod_{j=1}^{k} p_j^{\delta_j - 1}$$

where $p_j \in (0, 1)$ and $\sum_{j=1}^{k} p_j = 1$, $k \geq 2$, and $\delta_j > 0$.

Let $\mathbf{x} = (x_1, ..., x_n)$, then the likelihood density is:

$$\begin{aligned} f(x|p) &= \prod_{i=1}^{n} f(x_i|p) \\ &= \frac{\Gamma(\sum_{j=1}^{k} x_{ij} + 1)}{\prod_{j=1}^{k} \Gamma(x_{ij} + 1)} \prod_{j=1}^{k} p_j^{x_{ij}}. \end{aligned}$$

The posterior density is:

$$\begin{aligned} f(p|x) &\propto f(x|p)f(p) \\ &= \prod_{i=1}^{n} \left(\frac{\Gamma(\sum_{j=1}^{k} x_{ij} + 1)}{\prod_{j=1}^{k} \Gamma(x_{ij} + 1)} \prod_{j=1}^{k} p_j^{x_{ij}} \right) \frac{\Gamma(\sum_{j=1}^{k} \delta_j)}{\prod_{j=1}^{k} \Gamma(\delta_j)} \prod_{j=1}^{k} p_j^{\delta_j - 1} \\ &\propto \prod_{i=1}^{n} \prod_{j=1}^{k} p_j^{x_{ij}} \prod_{j=1}^{k} p_j^{\delta_j - 1} \\ &= \prod_{j=1}^{k} p_j^{\sum_{i=1}^{n} x_{ij}} \prod_{j=1}^{k} p_j^{\delta_j - 1} \\ &= \prod_{j=1}^{k} p_j^{\sum_{i=1}^{n} x_{ij} + \delta_j - 1}. \end{aligned}$$

Obviously this is the density of a Dirichlet $(\sum_{i=1}^{n} x_{i1} + \delta_1, ..., \sum_{i=1}^{n} x_{ik} + \delta_k)$. Note that the posterior density $f(p|x)$ is in the same family as the prior density $f(p)$ with different parameters.

Therefore $f(p)$ is conjugate prior for p.

For the parameters λ_j

For parameters λ_j, $j = 1, ..., k$ an inverse gamma density is often chosen as a prior. That is if $f(\lambda_j)$ denote the prior density of the j^{th} parameter of exponential mixture then

$$\lambda_j \sim \text{inverse gamma}(\alpha, \beta).$$

Let $X_1, X_2, ..., X_n$ be $i.i.d$ exponential(θ), and suppose the prior density as $inversegamma(\alpha, \beta)$, that is, the prior density is given by

$$f(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{-\alpha-1} e^{-\frac{\beta}{\theta}}, \quad \theta > 0, \alpha > 0, \beta > 0.$$

The likelihood density is:

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}} \\ &= (\frac{1}{\theta} e^{-\frac{x_1}{\theta}})(\frac{1}{\theta} e^{-\frac{x_2}{\theta}})...(\frac{1}{\theta} e^{-\frac{x_n}{\theta}}) \\ &= \underbrace{\frac{1}{\theta}.\frac{1}{\theta}.....\frac{1}{\theta}}_{\text{n-copies}} e^{-\frac{x_1}{\theta}}.e^{-\frac{x_2}{\theta}}.....e^{-\frac{x_n}{\theta}} \\ &= \frac{1}{\theta^n} e^{-\frac{(\sum_{i=1}^{n} x_i)}{\theta}}. \end{aligned}$$

The posterior density is:

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta) f(\theta) \\ &= \frac{1}{\theta^n} e^{-\frac{(\sum_{i=1}^{n} x_i)}{\theta}} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-\alpha-1} e^{-\frac{\beta}{\theta}} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\theta^n} \theta^{-\alpha-1} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}} e^{-\frac{\beta}{\theta}} \\ &= \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)}}_{\text{does not involve } \theta} \theta^{-n-\alpha-1} e^{-\frac{(\sum_{i=1}^{n} x_i+\beta)}{\theta}} \end{aligned}$$

The posterior density becomes:

$$f(\theta|x) \propto \theta^{-(n+\alpha)-1} e^{-\frac{(\sum_{i=1}^{n} x_i+\beta)}{\theta}}.$$

Clearly this is an inverse gamma density with parameters $n+\alpha$, and $\sum_{i=1}^{n} x_i + \beta$.

So,

$$(\theta|x) \sim \text{inverse gamma} \left(n+\alpha, \sum_{i=1}^{n} x_i + \beta \right)$$

Note that the posterior density $f(\theta|x)$ is in the same family as the prior density $f(\theta)$ with different parameters. Therefore $f(\theta)$ is conjugate prior for θ.

4.1.3 The posterior density

By using the conditional independence, the joint density of all variables can be written in general as

$$\begin{aligned} f(\lambda, p, z, x) &= f(x, z|\lambda, p) f(\lambda, p) \\ &= f(x, z|\lambda, p) g(\lambda)\pi(p) \quad (\ \lambda \text{ and } p \text{ are independent}) \end{aligned}$$

where $p = (p_j)_{j=1}^k$, $z = (z_{ij})_{j=1}^k$, $\lambda = (\lambda_j)_{j=1}^k$, $x = (x_i)_{i=1}^n$.

Note that parameters $\lambda_j s$ are independent, so the prior joint density for λ is then given by

$$g(\lambda) = g(\lambda_1)...g(\lambda_k).$$

By Bayes' theorem the posterior joint density given by

$$\begin{aligned} f(\lambda, p, z|x) &= \frac{f(\lambda, p, z, x)}{f(x)} \\ &\propto f(\lambda, p, z, x) \\ &= f(x, z|\lambda, p) g(\lambda)\pi(p) \\ &= f(x|\lambda, z) f(z|p) g(\lambda)\pi(p) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{1}{\lambda_j} e^{\frac{x_i}{\lambda_j}}\right)^{z_{ij}} \prod_{i=1}^{n}\prod_{j=1}^{k}(p_j)^{z_{ij}} \times g(\lambda_1) \times ... \times g(\lambda_k) \times (k-1)! \end{aligned}$$

where $g(\lambda_j) \sim inversegamma(\alpha, \beta)$, with different variables α, β for each λ_j, $j = 1, ..., k$.

4.2 Full Conditional Posterior Distributions

In this section we will find the full conditional distributions, using the Gibbs sampler method, which is one of a set of Markove chain Monto Carlo (MCMC) methods, in which the full conditional posterior distributions of all parameters are required. Using our likelihood, priors, and posterior joint density we

obtain all full conditional posterior densities by ignoring all terms that are constant with respect to the parameter.

Recall that for our finite exponential mixture, the likelihood distribution is:

$$f(x, z|\lambda, p) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left(p_j \frac{1}{\lambda_j} e^{-\frac{x_i}{\lambda_j}}\right)^{z_{ij}}$$

and our priors for λ_j, $j = 1, ..., k$, are

$$\lambda_j \sim \text{inverse gamma}(\alpha, \beta)$$

and the priors for proportions p,

$$p \sim \text{Dirichlet}(\delta, \delta, ..., \delta), \quad \delta = 1.$$

4.2.1 λ_j Posterior

The full conditional posterior density for λ_j is

$$\begin{aligned}
f(\lambda_j|\lambda_1, ..., \lambda_{j-1}, \lambda_{j+1}, ..., \lambda_k, p, z, x) &\propto f(x|\lambda, z)g(\lambda_j) \\
&= \left(\prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{1}{\lambda_j} e^{-\frac{x_i}{\lambda_j}}\right)^{z_{ij}}\right) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_j^{-\alpha-1} e^{-\frac{\beta}{\lambda_j}} \\
&= \prod_{i=1}^{n}\left(\frac{1}{\lambda_j} e^{-\frac{x_i}{\lambda_j}}\right)^{z_{ij}} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_j^{-\alpha-1} e^{-\frac{\beta}{\lambda_j}} \\
&\propto \prod_{i=1}^{n}\left(\frac{1}{\lambda_j} e^{-\frac{x_i}{\lambda_j}}\right)^{z_{ij}} \times \lambda_j^{-\alpha-1} e^{-\frac{\beta}{\lambda_j}} \\
&= \left(\frac{1}{\lambda_j} e^{-\frac{x_1}{\lambda_j}}\right)^{z_{1j}} \times ... \times \left(\frac{1}{\lambda_j} e^{-\frac{x_n}{\lambda_j}}\right)^{z_{nj}} \times \lambda_j^{-\alpha-1} e^{-\frac{\beta}{\lambda_j}} \\
&= \left(\frac{1}{\lambda_j}\right)^{z_{1j}} \times ... \times \left(\frac{1}{\lambda_j}\right)^{z_{nj}} \times e^{-\frac{x_1 z_{1j}}{\lambda_j}} \times ... \times e^{-\frac{x_n z_{nj}}{\lambda_j}} \\
&\times \lambda_j^{-\alpha-1} \times e^{-\frac{\beta}{\lambda_j}} \\
&= \left(\frac{1}{\lambda_j}\right)^{\sum_{i=1}^{n} z_{ij}} .\lambda_j^{-\alpha-1} .e^{-\frac{1}{\lambda_j}\sum_{i=1}^{n} x_i z_{ij}} .e^{-\frac{1}{\lambda_j}\beta} \\
&= \lambda_j^{-(\sum_{i=1}^{n} z_{ij}+\alpha)-1} .e^{-\frac{1}{\lambda_j}(\sum_{i=1}^{n} x_i z_{ij}+\beta)}
\end{aligned}$$

Therefore,

$$\lambda_j \sim \text{inverse gamma}(\sum_{i=1}^{n} z_{ij} + \alpha, \sum_{i=1}^{n} x_i z_{ij} + \beta) \tag{4.2}$$

Remark 4.2.1. The second relation in the previous equations because we ignore all terms that does not involve λ_j, so we omit p_j, and

$$\prod_{1}^{j-1} \left(\frac{1}{\lambda_j} e^{-\frac{x_i}{\lambda_j}} \right)^{z_{ij}} \prod_{j+1}^{k} \left(\frac{1}{\lambda_j} e^{-\frac{x_i}{\lambda_j}} \right)^{z_{ij}}$$

4.2.2 p Posterior

The full conditional posterior density for p is

$$\begin{aligned} f(p|z, \lambda, x) &\propto f(z|p)\pi(p) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} (p_{ij})^{z_{ij}} (k-1)! \\ &= \prod_{j=1}^{k} p_j^{\sum_{i=1}^{n} z_{ij}} (k-1)! \\ &\propto \prod_{j=1}^{k} p_j^{\sum_{i=1}^{n} z_{ij}} \\ &= \prod_{j=1}^{k} p_j^{(\sum_{i=1}^{n} z_{ij}+1)-1} \end{aligned}$$

So, from last equation we note that

$$p \sim \text{Dirichlet} \left(1 + \sum_{i=1}^{n} z_{i1}, ..., 1 + \sum_{i=1}^{n} z_{ik} \right). \tag{4.3}$$

4.2.3 z_i Posterior

For each observation x_i, i=1,...,n we have an indicator z_i such that

$$z_i = (z_{i1}, z_{i2}, ..., z_{ik}) = (z_{ij})_{j=1}^{k}$$

where each z_{ij} takes on two values only 1 or 0, and for each z_i only one of $z_{ij}^{'}s$ equal to 1 and the rest are all 0.
Therefore for fixed i, $\sum_{j=1}^{k} z_{ij} = 1$.
Using Bayes' theorem we have,
for fixed i, $i = 1, ..., n$, and for $j = 1, ..., k$,

$$\begin{aligned} f(z_{ij} = 1|x_i, \lambda, p) &= \frac{f(x_i|\lambda, p, z_{ij} = 1) f(z_{ij} = 1|\lambda, p, x_i)}{\sum_{j=1}^{k} f(x_i|\lambda, p, z_{ij} = 1) f(z_{ij} = 1|\lambda, p, x_i)} \\ &= \frac{f(x_i|\lambda, z_{ij} = 1) f(z_{ij} = 1|p)}{\sum_{j=1}^{k} f(x_i|\lambda, z_{ij} = 1) f(z_{ij} = 1|p)} \\ &= \frac{f(x_i|\lambda_j) p_j}{\sum_{j=1}^{k} f(x_i|\lambda_j) p_j} \\ &= \frac{f(x_i|\lambda_j) p_j}{f(x_i)}. \end{aligned}$$

Since each z_{ij} takes two values only 1 or 0, then

$$f(z_{ij} = 0|x_i, \lambda, p) = 1 - f(z_{ij} = 1|x_i, \lambda, p) = 1 - \frac{f(x_i|\lambda_j) p_j}{f(x_i)}.$$

Thus

$$z_{ij} \sim \text{Bernoulli} \left(\frac{f(x_i|\lambda_j) p_j}{f(x_i)} \right)$$

so $z_i = (z_{ij})_{j=1}^{k} \sim$ multinomial $(1, w_{i1}, ..., w_{ik})$, $i = 1, ..., n$, $j = 1, ..., k$, where

$$w_{ij} = \frac{f(x_i|\lambda_j) p_j}{f(x_i)}, \quad j = 1, ..., k. \quad \text{See [5].}$$

4.2.4 Gibbs Updates for Fixed k

We consider a mixture of exponentials where, conditional on there being k components in the mixture. All the parameters of our exponential mixture have full conditional densities that are well known and easy to sample from. We can therefore perform Gibbs updates on them where the draws are from their full conditionals. The general Gibbs algorithm for fixed k is then

Step 1: Pick a starting values of the parameters for the Markov chain, say $(\lambda_1^0, ..., \lambda_k^0, p^0, z_1^0, ..., z_n^0)$

Step 2: Update each variable in turn at the l^{th} iteration, $l = 1, ..., N$:

(a) Gibbs update of λ_j: $j = 1, ..., k$: Sample λ_j^l from gamma $(n + \alpha_j, \sum_{i=1}^n z_{ij} + \beta_j)$ using the most up-to-date values of z_{ij} .

(b) Gibbs update of proportions p: Sample p^l from Dirichlet $(1 + \sum_{i=1}^n z_{i1}, ..., 1 + \sum_{i=1}^n z_{ik})$ using the most up-to-date values of $z_{i1}, ..., z_{ik}$.

(c) Gibbs update of indicators z_i: Sample z_i^l from multinomial $(1, w_{i1}, ..., w_{ik})$ $i = 1, ..., n$, $j = 1, ..., k$, where

$$w_{ij} = \frac{f(x_i|\lambda_j)p_j}{f(x_i)} \qquad , j = 1, ..., k.$$

using the most up-to-date values of λ_j and p.

(d) We now have a new Markov chain state $(\lambda_1^l, ..., \lambda_k^l, p^l, z_1^l, z_k^l)$

Step 3: Return to **step 2**, $N - 1$ times to produce a Markov chain of length N. See[12]

4.3 Study Case

In this section we apply our exponential mixture model of two components on real data example to illustrate our methodology.

Our example uses a dataset from length of hospital stay project (LOS). This data study the length stay of patients in a psychiatric hospital in North East London in 1991 and this was studied by Harrison and Millard (1991) and McClean and Millard (1993).

Health service researchers frequently study length of hospital stay (LOS) as a health outcome. Generally originating from heavily skewed distributions,

LOS data can be difficult to model with a single parametric model. Mixture models can be quite effective in dealing with such data. This example illustrates how to perform a Bayesian analysis of an exponential mixture model for LOS data. The experimental MCMC procedure is used for this analysis. See[42]

We simulate a sample using the Gibbs sampler which use the full conditional distributions derived in the previous section. We do this by using an R script that we modify to suit our exponential mixture model of two components, we employ it to generate samples to make estimation of the unknown parameters of the model, and to perform the required Bayesian analysis by using the simulation results.

4.3.1 Estimation results

We choose the initial values for the parameters α, β, p_1, p_2, λ_1, and λ_2 like this:

$$\alpha = \beta = 0.5$$

$$p_1 = p_2 = 0.5$$

$$\lambda_1 = 600, \quad \lambda_2 = 5000$$

Summary for λ_1:

Table 4.1: Summary for λ_1

Min	1st Qu	median	mean	3rd Qu	Max	sd
353.9	571.6	619.8	621.2	670.1	924.8	73.7254

Summary for λ_2:

Table 4.2: Summary for λ_2

Min	1st Qu	median	mean	3rd Qu	Max	sd
5000	7283	7733	7786	8233	12090	729.31

Summary for data:

Min	1st Qu	median	mean	3rd Qu	Max
1	285	1134	3712	3855	24030

- p_1: Mean= 0.5648608, sd=0.03947177
- p_2: Mean= 0.4351392, sd=0.03947177
- z_1: Mean= 0.5651891, sd=0.4957322
- z_2: Mean= 0.4349109, sd=0.4957454

4.3.2 Simulation results

These figures show that the data is not homogenous, so we don't use the usual homogenous exponential distribution and we use mixture model to express this data.

From these figures for λ_1 we see that λ_1 is symmetric and has gamma density.

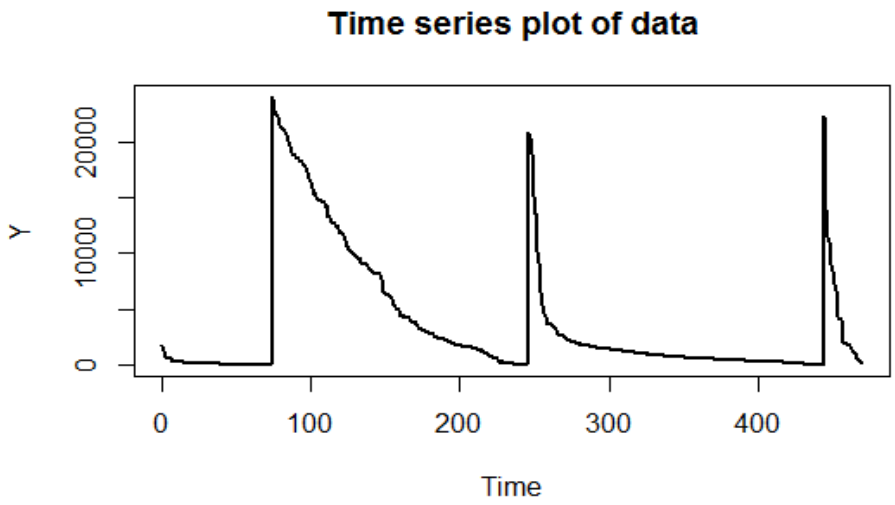

Figure 4.1: Time series plot for the data.

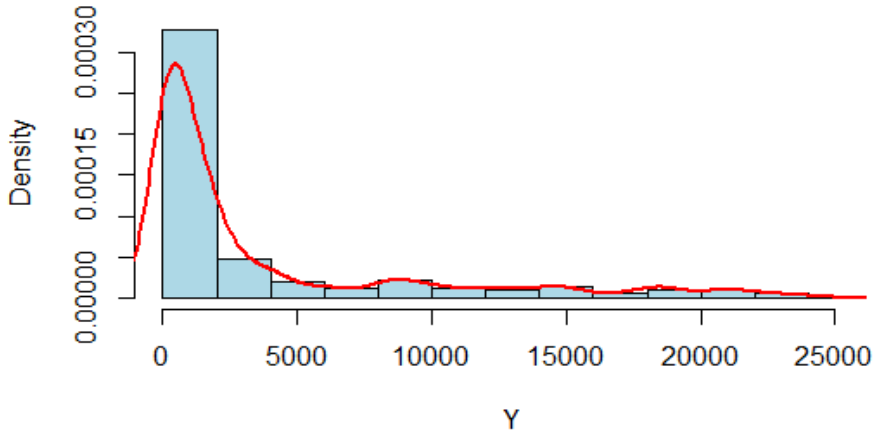

Figure 4.2: Data plot and its histogram.

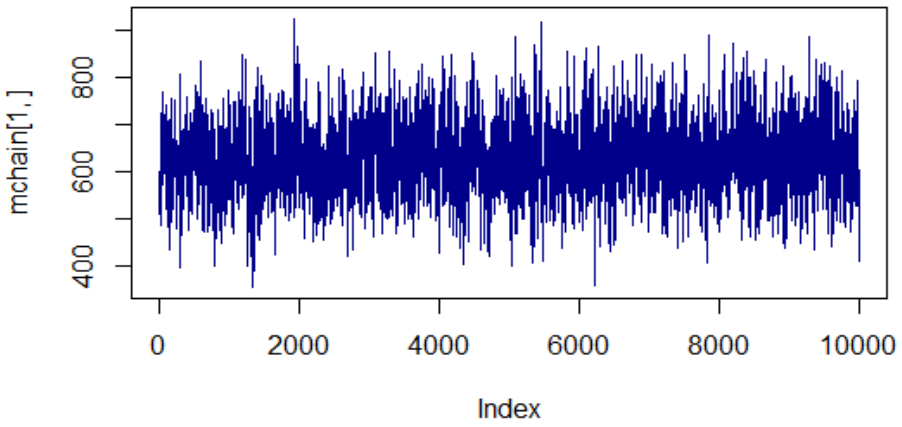

Figure 4.3: Markov chain for λ_1.

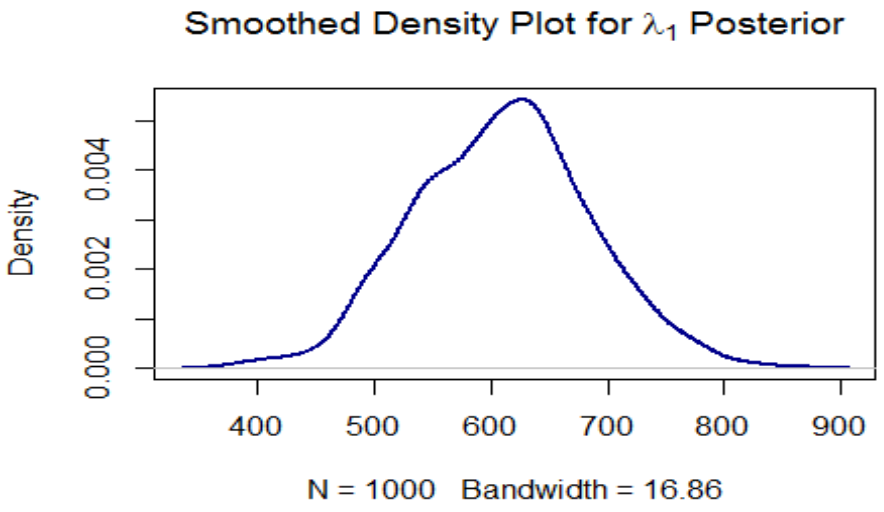

Figure 4.4: Density plot for λ_1.

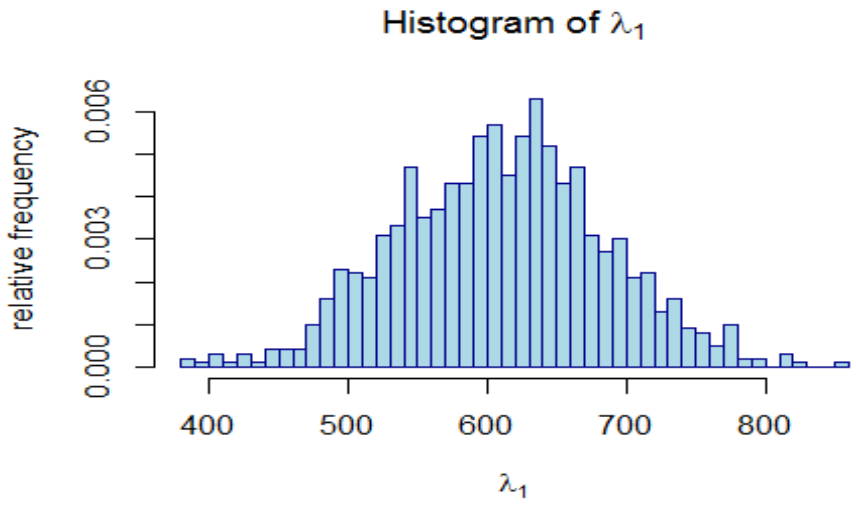

Figure 4.5: Histogram for λ_1.

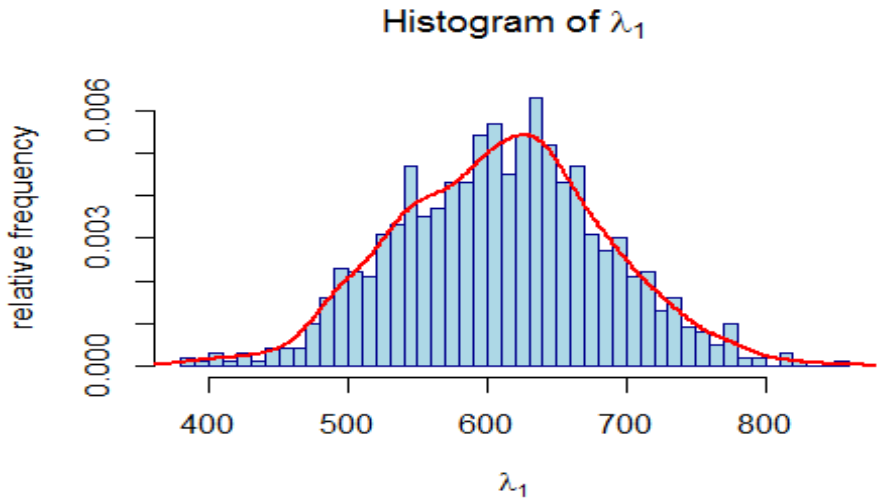

Figure 4.6: Density plot for λ_1 and its histogram.

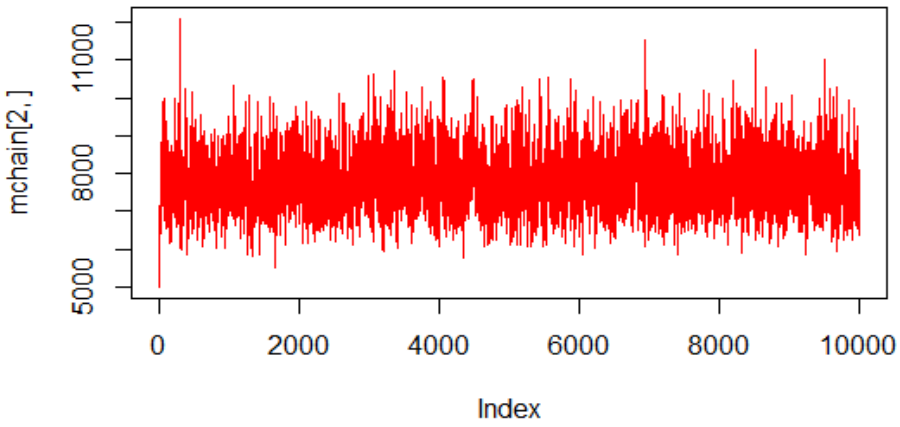

Figure 4.7: Markov chain for λ_2.

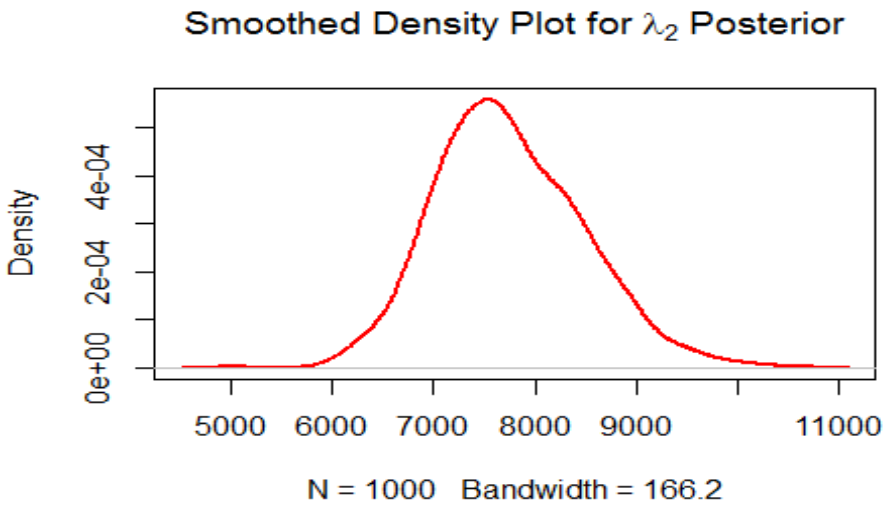

Figure 4.8: Density plot for λ_2.

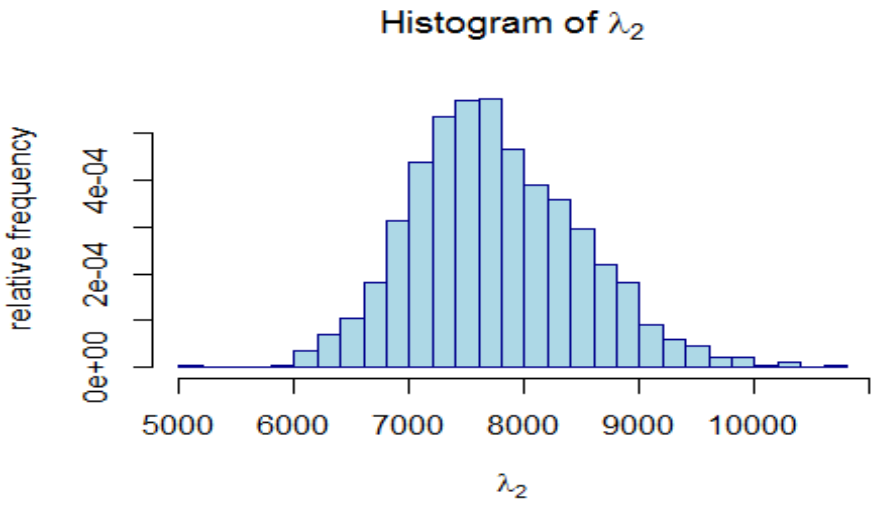

Figure 4.9: Histogram for λ_2.

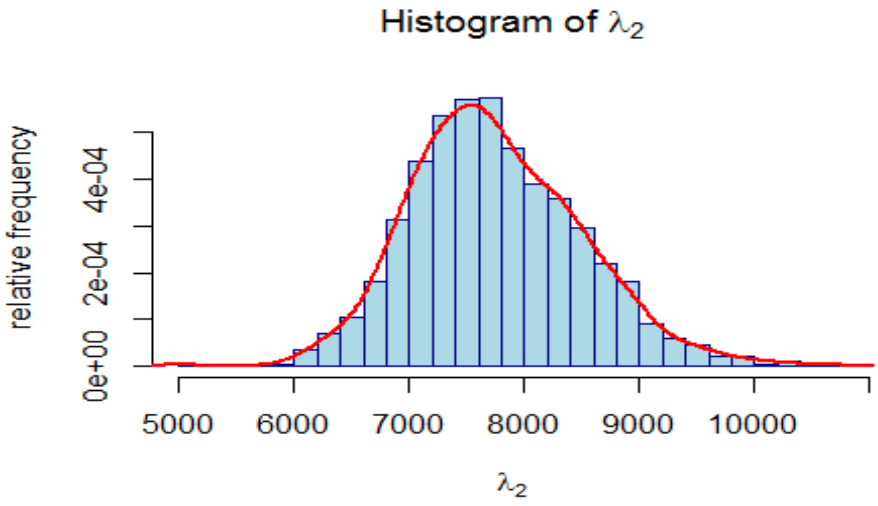

Figure 4.10: Density plot for λ_2 and its histogram.

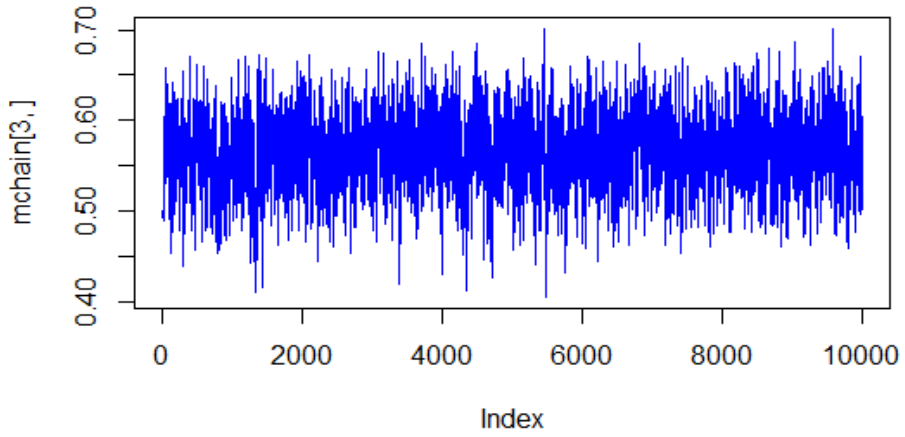

Figure 4.11: Markov chain for p_1.

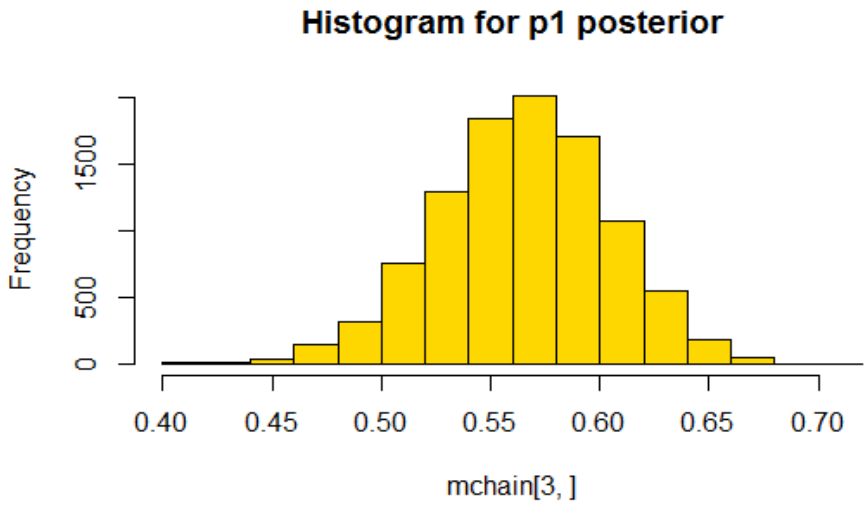

Figure 4.12: Histogram for p_1.

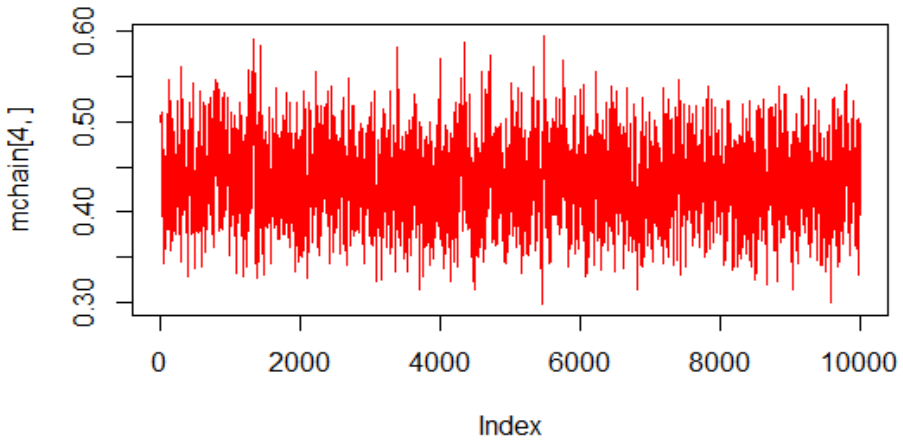

Figure 4.13: Markov chain for p_2.

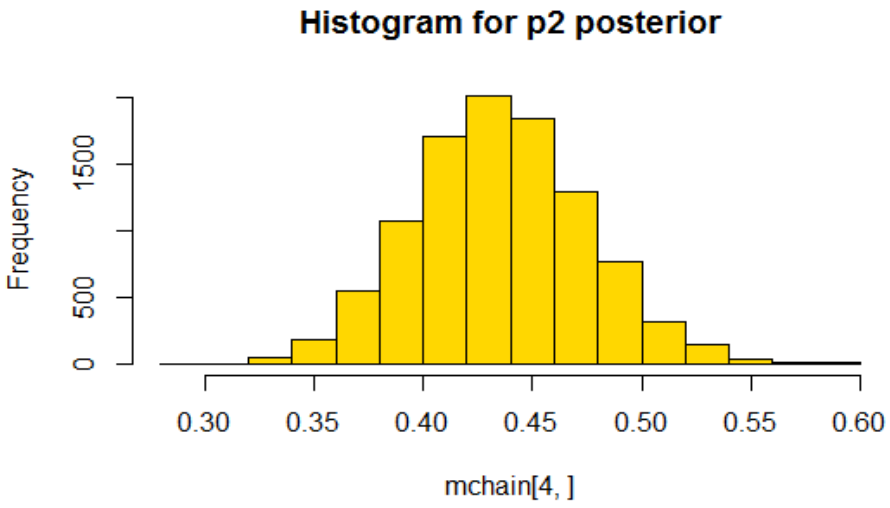

Figure 4.14: Histogram for p_2.

Bibliography

[1] Barut, A. E. (2010). *Mixture Models And E-M Algorithm*, Princeton University, New Jersey.

[2] Bolstad, W. M. (2010). *Understanding Computational Bayesian Statistics*, John Wiley & Sons, Inc., New Jersey.

[3] Brown, G. O., Brooks, S. P. and Buckley, W. S. (2010). *Experience Rating with Poisson Mixtures*, Centre for Mathematical Sciences, Cambridge.

[4] Ching, W.-K. and Michael, K. N. (2006). *Markov Chains Models, Algorithms and Applications*, Springer Science+Business Media, Inc, USA.

[5] Dellaportas, P., Karlis, D. and Xekalaki, E. (1997). *Bayesian Analysis of Finite Poisson Mixtures*, Athens University of Economics and Business, Greece.

[6] Durrett, R.(2012). *Essentials of Stochastic Processes*, Springer Texts in Statistics, New York, second edition.

[7] Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*, Cambridge University Press, Cambridge.

[8] Gelman, A., Carlin J. B. Stern, H. S. and Rubin, D. B. (2009). *Bayesian Data Analysis*, CRC Press, Florida, second edition.

[9] Ghahramani, S. (2005). *Fundamentals of Probability with Stochastic Processes*, Upper Saddle River, New Jersey.

[10] Ghosh, J. K. (2006). *An Introduction to Stochastic Processes*, Springer, USA.

[11] Häggström, O. (2003). *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press, Cambridge.

[12] Haran, M. (2014). *Bayesian Change Point Model With Gamma Hyperpriors*, Penn State University, Pennsylvania.

[13] Hoff, P. D. (2006). *Introduction to Bayesian Statistics for the Social Sciences*, University of Washington, Washington.

[14] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*, Springer Science+Business Media, LLC., USA.

[15] Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*, John Wiley & Sons, UK.

[16] Johnson, R. A. (2010). *Bayesian Inference*, General Books LLC, Madison.

[17] Lee, K., Marin, J.-M., Mengersen, K. and Robert, C. (2008). *Bayesian Inference on Mixtures of Distributions*, Platinum Jubilee of the Indian Statistical Institute, Bangalore.

[18] Leveque, O. (2011). *Lecture Notes on Markov Chains*, National University of Ireland, Maynooth.

[19] Lin, M.-Y. (2013). *Bayesian Statistics*, Boston University, Boston.

[20] Marin, J.-M., Mengersen, K. and Robert, C. P. (2005). *Bayesian Modelling and Inference on Mixtures of Distributions*, Handbook of Statistics, 459-507, North Holland.

[21] Mclachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, Inc., USA.

[22] Mengersen, K. L., Robert, C. P. and Titterington, D. M. (2011). *Mixtures Estimation and Applications*, John Wiley & Sons, Ltd, United Kingdom.

[23] Michigan, W. (2008). *Discrete-Time Markov Chains*, The University of Hong Kong.

[24] Nobile, A. (1994). *Bayesian Analysis of Finite Mixture Distribution*, Carnegie Mellon University, Pennsylvania.

[25] Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*, Springer Science+Business Media, New York.

[26] Rufo, M. J., Martin, J. and Perez, C. J. (2006). *Bayesian analysis of finite mixture models of distributions from exponential families*, University of Extremadura, Spain.

[27] Sahoo, P. (2013). *Probability And Mathematical Statistics*, University of Louisville, USA.

[28] Sahu, S. K. (2001). *Bayesian Methods*, University of Southampton, UK.

[29] Semieniuk, G. and Scharfenaker, E. (2014). *A Bayesian Latent Variable Mixture Model for Filtering Firm Profit Rates*, The New School for Social Research, New York.

[30] Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*, Magdalen College, Oxford.

[31] Teicher H. (1963). *Identifiability of Finite Mixtures*, Annals of Mathematical Statistics.

[32] Uysal, D. (2012). *Properties of a Random Sample*, IHS, Vienna.

[33] Viallefont, V., Richardson, S. and Green, P. J. (2002). *Bayesian Analysis of Poisson Mixtures*, Journal of Nonparametric Statistics 14, 181-202.

[34] Weber, R. (2011). *Markov Chains*, Cambridge University, Cambridge.

[35] Wilkinson, D. (1998). *Introduction to Probability and Statistics*, School of Mathematics & Statistics, London.

[36] Jackman, S. (2009). *Bayesian Modeling in the Social Science*, Library of Congress Cataloguing, India.

[37] Zhong, J. (2012). *The Diagnostic for Poisson Mixture and Application*, Shanghai University of Finance and Economics, China.

[38] http://www.astro.cornell.edu/staff/loredo/bayes, Jan. 25, 2016.

[39] https://en.wikipedia.org/wiki/Dirichlet_distribution, Oct. 11, 2015.

[40] https://en.wikipedia.org/wiki/Multinomial_distribution, Oct. 11, 2015.

[41] http://sites.stat.psu.edu/ mharan/MCMCtut/MCMC.html, Nov. 7, 2015.

[42] http://support.sas.com/rnd/app/examples/stat/BayesMixtureExp/new example, Jan. 27, 2016.