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

Bayesian Inference on Finite Mixtures of Poisson Distributions

Presented by
Duaa F. Abu Hamdah

Supervisor
Prof. Mohamed I. Riffi

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master in Mathematics
2015

Dedication

This thesis is dedicated
firstly to my parents. I have to thank you for your endless love, support, and encouragement.

To my husband, Ali, you have been a constant source of support and motivation during the challenges of graduate and life. You have been patient with me when I am frustrated, you celebrate with me when even the littlest things go right, and you are there whenever I need you to just listen. I am truly thankful for you for having you in my life.

To my daughters Deema and Alma and my son Anas, who shared me in every step of success in my life.

To my sisters, brother, deserve my wholehearted thanks as well.

To all my friends, thank you for your understanding and encouragement. Your friendship makes my life a wonderful experience.

Acknowledgements

I would first like to thank my supervisor, Prof. Mohamed I. Riffi. I am grateful to him for his guidance and the opportunities he has afforded me. He is incredibly organized and a great problem solver, both of these qualities were immensely helpful in moving my thesis forward.

I would also like to thank my thesis committee members, Dr. Esam Mahdi and Dr. Hazem El Shekh Ahmed, for their contributions to this work.

الملخص

توزيعات بواسون المختلطة تستخدم على نطاق واسع في مختلف التخصصات لتمثل البيانات في حالة أن كل قيمة ملاحظة يفترض ان تأتي من واحدة من عدة توزيعات بواسون مع معلمات مختلفة.

في هذه الأطروحة، ندرس الاستدلال البايزي لنموذج مزيج بواسون محدود باستخدام معاين جيبس باعتباره واحد من أهم طرق سلسة ماركوف مونتي كارلو.

نهجنا في هذه الأطروحة يعتمد على استخدام معاين جيبس لمحاكاة سلسلة ماركوف بحيث ان توزيع الكثافة الخلفي يمثل توزيعها الثابت.

ثم نستخدم العينة الناتجة لعمل حسابات الاستدلال البايزي و عمل الاستناتاجات حول المعلمات المجهولة لمزيج بواسون. و نختتم هذه الأطروحة بتقديم مثال لبيانات حقيقية لتوضيح المنهجية التي اتبعناها.

Abstract

Mixed Poisson distributions are widely used in various disciplines to model data in which each observation is assumed to come from one of a number of Poisson distributions with different parameters. In this thesis, we investigate the Bayesian estimation for the finite Poisson 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 stationary distribution. Then we use the resulting sample to make the suitable Bayesian computations and draw conclusion about the unknown parameters of the Poisson mixture model. We conclude this thesis by presenting a real data example to illustrate our methodology.

Contents

List of Tables

List of Figures

List of Abbreviations

gcd	greatest common divisor.
iid	independent identically distributed.
MC	Markov Chain.
MCMC	Markov Chain Monte Carlo.
M-H	Metropolis-Hastings.
sd	standard deviation.
pdf	probability density function.
pmf	probability mass function.

Introduction

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

There are situations where the simple Poisson distribution model becomes inadequate to model a data that contains a large amount of over dispersion. We face this challenge by using the Poisson 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 comprising 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. 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.

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 Poisson mixtures in applications. Then we mention the approach we are going to follow in making Bayesian inference about the Poisson 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 Poisson mixtures model using the missing data formulation.

Chapter 3 Markov Chains

In this chapter, we give a brief introduction on discrete-time Markov chains also, we will discuss some basic properties of a Markov chain. Basic concepts and notations are explained also some important theorems in this area will be presented.

Chapter 4 Markov Chain Monte Carlo Methods

In this chapter we look at Markov chain Monte Carlo (MCMC) methods for generating samples from the posterior distribution. We present the Gibbs sampler and algorithm as one of the most basic Markov Chain Monte Carlo (MCMC) methods in Bayesian analysis. Also, we present the algorithm used to generate samples.

Chapter 5 Bayesian Analysis of Finite Poisson Mixtures

In this chapter we use the Gibbs sampler and algorithm to draw samples from the posterior of the Poisson 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

In this chapter we present an introduction to Bayesian Statistics including the following topics: Bayes theorem, model-based Bayesian inference, expressing the posterior probability density function in terms of the prior density and the likelihood function, conjugate priors, posterior predictive distributions, and some related examples. Finally we compare the frequentist approach with the Bayesian approach and we show advantages of Bayesian inference over frequentist 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 is a very clever restatement of the conditional probability formula. It gives a method for updating the probabilities of unobserved events, given that another related event has occurred. This means that we have a prior probability for the unobserved event, and we update this to get its posterior probability, given the occurrence of the related event. In Bayesian statistics, Bayes theorem is used as the basis for inference about the unknown parameters of a statistical distribution.

Since we are uncertain about the true values of the parameters, in Bayesian statistics we will consider them to be random variables. This contrasts with the non-Bayesian statistics that the parameters are fixed but unknown constants.

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].

1.2 Bayes Theorem

In this section we give some basics of probability theory that will be needed from now on, including the most important part, which is Bayes theorem.

Definition 1.2.1. [28](Conditional Probability)

Let A and B be events from a given event space F with B satisfying $P(B) > 0$.

The conditional probability of A, given that B occurs, is a probability measure denoted by $P(A|B)$ and is defined by,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.1}$$

If $P(B) = 0$ then $P(A|B)$ is not defined.

Proposition 1.2.1. [36](Multiplication Rule)

Let A and B be two events with $P(A) > 0$, and $P(B) > 0$. Then

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \tag{1.2}$$

Definition 1.2.2. [36] (Independence)

Two events, A and B with $P(A) > 0$, and $P(B) > 0$, are said to be independent when,

$$P(A \cap B) = P(A)P(B).$$

Or equivalently when,

$$P(A|B) = P(A).$$

Also, A and B are said to be conditionally independent, given C, when

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Definition 1.2.3. [28] (Partition)

A collection of events $B_1, B_2, ..., B_k$ is called a partition of the sample space Ω if

1. $B_i \cap B_j = \phi$, for all i and j such that $i \neq j$,

2. $B_1 \cup B_2 \cup ... \cup B_k = \Omega$.

Theorem 1.2.1. *[28] (Total Probability Theorem)*

For any partition, $B_1, B_2, B_3, ..., B_k$ of the sample space with $P(B_j) > 0$ for all $1 \leq j \leq k$, we have

$$P(A) = \sum_{j=1}^{k} P(A|B_j)P(B_j). \tag{1.3}$$

Theorem 1.2.2. *[28] (Bayes' Theorem (A special case))*

For any events A, and B with $P(A) > 0$ and $P(B) > 0$,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}. \tag{1.4}$$

Theorem 1.2.3. *[28] (Bayes' Theorem)*

Let A be an event with $P(A) > 0$. Let $B_1, B_2, ..., B_n$ form a partition of Ω such that $P(B_i) > 0$ for all $1 \leq i \leq n$.

Then, for each $j = 1, 2, ..., n$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}. \tag{1.5}$$

Theorem 1.2.4. *[15](Bayes Theorem for Continuous Parameters)*

Suppose that two continuous random variables X and θ are given with pdfs $f(x|\theta)$ and $f(\theta)$. Then

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)f(\theta)d\theta}. \tag{1.6}$$

Definition 1.2.4. [33] The random variables $X_1, ..., X_n$ are called **a random sample** of size n from the distribution $f(x)$, if $X_1, ..., X_n$ are mutually independent random variables, and the marginal *pdf* or *pmf* of each X_i is the same probability distribution as the others. Alternatively, $X_1, ..., X_n$ are called **independent and identically distributed** random variables with *pdf* or *pmf* $f(x)$. This is

commonly abbreviated as **iid random variables**.

If the population *pdf* or *pmf* is a member of a parametric family with *pdf* or *pmf* given by $f(x|\theta)$, then the joint *pdf* or *pmf* is

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where the same parameter value θ is used in each of the terms in the product.

Definition 1.2.5. [23] Given iid random sample $\{Y_i : i = 1, \ldots, n\}$ with a density in the parametric family $\{f(y_i|\theta) : i = 1, ..., n\}$, one can construct the likelihood density :

$$f(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

1.3 Model-Based Bayesian Inference

The basis for Bayesian inference is derived from Bayes' theorem. Here is Bayes' theorem, equation 1.4, again

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

replacing B with observations $\mathbf{x} = (x_1, x_2, ..., x_n)$, A with a parameter θ, and probabilities P with a function f, results in the following

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})} \tag{1.7}$$

where $f(\mathbf{x})$ will be discussed below.

$f(\theta)$ is the prior distribution of parameter θ before y is observed,

$f(\mathbf{x}|\theta)$ is the likelihood of $\mathbf{x}$ under a model,

and $f(\theta|\mathbf{x})$ is the posterior distribution of the parameter θ. See[16].

Note 1.3.1. Since there are usually multiple parameters, we can replace θ in equation 1.7 by Θ to represent a set of j parameters, and may be considered as $\Theta = (\theta_1, \theta_2, ..., \theta_j)$.

Now to discuss the denominator $f(\mathbf{x})$ we need the following definition.

Definition 1.3.1. [29] (Marginal likelihood or the prior predictive distribution)

Let $X_1, X_2, ..., X_n$ be an iid sample from the distribution $f(\mathbf{x}|\theta)$. Let $f(\theta)$ be the prior distribution and $f(\theta|\mathbf{x})$ be the posterior distribution.

Then the marginal likelihood or the prior predictive distribution is given by:

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta) f(\theta) d\theta. \tag{1.8}$$

So the denominator $f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta) f(\theta) d\theta$ defines the marginal likelihood of $\mathbf{x}$, or the prior predictive distribution of $\mathbf{x}$, and may be set to an unknown constant c.
The prior predictive distribution indicates what $\mathbf{x}$ should look like, given the model, before $\mathbf{x}$ has been observed. Only the set of prior probabilities and the model's likelihood function are used for the marginal likelihood of $\mathbf{x}$.

Remark 1.3.1. In probability theory, a normalizing constant is a constant by which an everywhere non-negative function must be multiplied so the area under its graph is 1, e.g., to make it a probability density function or a probability mass function.
For example, if we define

$$f(x) = e^{-x^2/2},\ x \in (-\infty, \infty)$$

we have

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}$$

if we define function $\varphi(x)$ as

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

so that

$$\int_{-\infty}^{\infty} \varphi(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1$$

Function $\varphi(x)$ is a probability density function. This is the density of the standard normal distribution. And constant $\dfrac{1}{\sqrt{2\pi}}$ is the normalizing constant of function $f(x)$.

By the previous remark and return to Equations 1.7, 1.8 we find that, the presence of the marginal likelihood of $\mathbf{x}$ normalizes the joint posterior distribution, $f(\theta|\mathbf{x})$, ensuring it is a probability distribution and integrates to one.
By replacing $f(\mathbf{x})$ with c, which is short for a constant of proportionality which is usually called the normalizing constant of function $f(\mathbf{x})$, the model-based formulation of Bayes' theorem becomes

$$f(\theta|\mathrm{x}) = \frac{f(\mathbf{x}|\theta) f(\theta)}{c}. \tag{1.9}$$

By removing c from the equation, the relationship changes from 'equals' (=) to 'proportional to' ($\propto$).

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta) \tag{1.10}$$

in words:

Posterior $\propto$ Likelihood $\times$ Prior.

This form can be stated as the unnormalized posterior being proportional to the likelihood times the prior. See[23].

Example 1.3.1. Let $X_1, X_2, ..., X_n$ be iid Poisson(θ).
Suppose the prior density is given by:

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

The likelihood density is:

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

The posterior density is:

$$\begin{aligned}
f(\theta|\mathbf{x}) &\propto f(x|\theta)f(\theta) \\
&= \frac{1}{x_1!x_2!...x_n!} e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i} e^{-\theta} \\
&= \frac{1}{x_1!x_2!...x_n!} e^{-n\theta-\theta}\theta^{\sum_{i=1}^{n} x_i}
\end{aligned}$$

$$= \underbrace{\frac{1}{x_1!x_2!...x_n!}}_{\text{does not involve } \theta} e^{-(n+1)\theta}\theta^{\sum_{i=1}^n x_i}.$$

We do not write the term which does not involve θ.

The posterior density becomes:

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

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

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

1.4 Conjugate Priors, Posterior Predictive Distributions

In this section, we present the definition of conjugate priors and the posterior predictive distribution and give some examples.

1.4.1 Conjugate Priors

Definition 1.4.1. [19] A family of probability distributions is said to be a conjugate prior family for iid sampling from a likelihood $f(\mathbf{x}|\theta)$, if, whenever the prior distribution for θ is a member of the family, then the posterior distribution for θ is also a member of that family, for any sample size and sample values.

Conjugate priors may not exist; when they do, selecting a member of the conjugate family as a prior is done mostly for mathematical convenience, since the posterior can be evaluated very simply. Table 1.1 provides some conjugate priors.

Table 1.1: Conjugate priors

Observations	Prior	Posterior
Poisson	Gamma	Gamma
Bernoulli	Beta	Beta
Binomial	Beta	Beta
Normal	Normal	Normal
Exponential	Gamma	Gamma

In the next example we prove the first case of Table 1.1.

Example 1.4.1. [29] Suppose $X_1, X_2, ..., X_n$ be iid $Poisson(\theta)$, and suppose the prior distributed as a gamma distribution that is, the prior density is given by

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

The likelihood density is:

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

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

The posterior density is:

$$
\begin{aligned}
f(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta) f(\theta) \\
&= \frac{1}{x_1! x_2! ... x_n!} e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha - 1} e^{-\beta\theta} \\
&= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{1}{x_1! x_2! ... x_n!} \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-n\theta - \beta\theta} \\
&= \underbrace{\frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{1}{x_1! x_2! ... x_n!}}_{\text{does not involve } \theta} \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-(n+\beta)\theta}.
\end{aligned}
$$

We do not write the term which does not involve θ.

The posterior density becomes:

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

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

So,

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

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

1.4.2 Posterior Predictive Distributions

Definition 1.4.2. [10] Let $X_1, X_2, ..., X_n$ be iid sample from the distribution $f(\mathbf{x}|\theta)$. Let $f(\theta)$ be the prior distribution and $f(\theta|\mathbf{x})$ be the posterior distribution. The posterior predictive distribution is

given by:

$$f(x_{n+1}|\mathbf{x}) = \int_{-\infty}^{\infty} f(x_{n+1}|\theta) f(\theta|\mathbf{x}) d\theta \tag{1.11}$$

Example 1.4.2. Suppose

$X_1, X_2, ..., X_n$ be iid $Poisson(\theta)$,

$\theta \sim gamma(\alpha, \beta)$.

Returning to Example 1.4.1, The likelihood density is:

$$f(\mathbf{x}|\theta) = \frac{1}{x_1! x_2! ... x_n!} e^{-n\theta} \theta^{\sum_{i=1}^n x_i}.$$

And the posterior density is:

$$(\theta|\mathbf{x}) \sim Ga(\sum_{i=1}^n x_i + \alpha, n + \beta).$$

Hence,

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

Here,

$$f(x_{n+1}) = \frac{\theta^{x_{n+1}} e^{-\theta}}{x_{n+1}!}.$$

The posterior predictive density is:

$$\begin{aligned} f(x_{n+1}|\mathbf{x}) &= \int_{-\infty}^{\infty} f(x_{n+1})|\theta) f(\theta|\mathbf{x}) d\theta \\ &= \int_0^{\infty} \frac{\theta^{x_{n+1}} e^{-\theta}}{x_{n+1}!} \frac{(n+\beta)^{\sum_{i=1}^n x_i + \alpha}}{\Gamma(\sum_{i=1}^n x_i + \alpha)} \theta^{\sum_{i=1}^n x_i + \alpha - 1} e^{(-n+\beta)\theta} d\theta \\ &= \frac{(n+\beta)^{\sum_{i=1}^n x_i + \alpha}}{\Gamma(\sum_{i=1}^n x_i + \alpha) x_{n+1}!} \int_0^{\infty} \theta^{\sum_{i=1}^{n+1} x_i + \alpha - 1} e^{(-n+\beta+1)\theta} d\theta. \end{aligned}$$

This integrand looks like a gamma distribution with parameters: $\sum_{i=1}^n x_i + \alpha$ and $n + \beta + 1$.

But,

$$\int_0^{\infty} \frac{(n+\beta+1)^{\sum_{i=1}^{n+1} x_i + \alpha}}{\Gamma(\sum_{i=1}^{n+1} x_i + \alpha)} \theta^{\sum_{i=1}^{n+1} x_i + \alpha - 1} e^{(-n+\beta+1)\theta} d\theta = 1$$

then,

$$\int_0^{\infty} \theta^{\sum_{i=1}^{n+1} x_i + \alpha - 1} e^{(-n+\beta+1)\theta} d\theta = \frac{\Gamma(\sum_{i=1}^{n+1} x_i + \alpha)}{(n+\beta+1)^{\sum_{i=1}^{n+1} x_i + \alpha}}.$$

Therefore

$$f(x_{n+1}|\mathbf{x}) = \frac{(n+\beta)^{\sum_{i=1}^n x_i + \alpha} \Gamma(\sum_{i=1}^{n+1} x_i + \alpha)}{\Gamma(\sum_{i=1}^n x_i + \alpha)(x_{n+1}!)(n+\beta+1)^{\sum_{i=1}^{n+1} x_i + \alpha}}.$$

Since

$$\Gamma(\sum_{i=1}^{n} x_i + \alpha) = (\sum_{i=1}^{n} x_i + \alpha - 1)!,$$
$$\Gamma(\sum_{i=1}^{n+1} x_i + \alpha) = (\sum_{i=1}^{n+1} x_i + \alpha - 1)!,$$

then,

$$f(x_{n+1}|\mathbf{x}) = \frac{(n+\beta)^{\sum_{i=1}^{n} x_i + \alpha}(\sum_{i=1}^{n+1} x_i + \alpha - 1)!}{(\sum_{i=1}^{n} x_i + \alpha - 1)!(x_{n+1}!)(n+\beta+1)^{\sum_{i=1}^{n+1} x_i + \alpha}}.$$

And since,

$$\frac{(\sum_{i=1}^{n+1} x_i + \alpha - 1)!}{(\sum_{i=1}^{n} x_i + \alpha - 1)!(x_{n+1})!} = \binom{\sum_{i=1}^{n+1} x_i + \alpha - 1}{x_{n+1}}$$

then,

$$f(x_{n+1}|\mathbf{x}) = \frac{(\sum_{i=1}^{n+1} x_i + \alpha - 1)!}{(\sum_{i=1}^{n} x_i + \alpha - 1)!(x_{n+1})!}\left(\frac{n+\beta}{n+\beta+1}\right)^{\sum_{i=1}^{n} x_i + \alpha}\left(\frac{1}{n+\beta+1}\right)^{x_{n+1}}$$

$$= \binom{\sum_{i=1}^{n+1} x_i + \alpha - 1}{x_{n+1}}\left(\frac{n+\beta}{n+\beta+1}\right)^{\sum_{i=1}^{n} x_i + \alpha}\left(\frac{1}{n+\beta+1}\right)^{x_{n+1}}.$$

Clearly this is the density of a negative-binomial distribution with parameters $\sum_{i=1}^{n+1} x_i + \alpha$ and $\frac{n+\beta}{n+\beta+1}$. That is

$$f(x_{n+1}|\mathbf{x}) \sim \textit{negative binomial}\left(\sum_{i=1}^{n+1} x_i + \alpha, \frac{n+\beta}{n+\beta+1}\right).$$

1.5 Difference Between Frequentist and Bayesian

In statistical inference, there are two broad categories of interpretations of probability: **Bayesian inference and frequentist inference**.

These views often differ with each other on the fundamental nature of probability. **Frequentist inference** loosely defines probability as the limit of an event's relative frequency in a large number of trials, and only in the context of experiments that are random and well-defined.

Bayesian inference, on the other hand, is able to assign probabilities to any statement, even when a random process is not involved. In Bayesian inference, probability is a way to represent an individual's degree of belief in a statement, or given evidence.

Within Bayesian inference, there are also different interpretations of probability, and different approaches based on those interpretations.

In recent years, Bayesian approach has been widely applied to clinical trials, research in education and psychology, and decision analyses. However, some statisticians still consider it as an interesting alternative to the classical theory based on relative frequency.

The following table briefly summarizes the differences between frequentist and Bayesian approaches. See [19].

Table 1.2: The differences between frequentist and Bayesian approaches.

	Frequentist	**Bayesian**
parameter of the model	• fixed, unknown constants • can NOT make probabilistic statements about the parameters	• random variables (parameters can't be determined exactly, uncertainty is expressed in probability statements or distributions) • can make probability statements about the parameters
probability	objective, relative frequency	subjective, degree of belief
main outcomes	point estimates with standard error	posterior distribution

	Frequentist	**Bayesian**
estimate/inference	use data to best estimate unknown parameters	• pinpoint a value of parameter space as well as possible by using data to update belief • all inference follow posterior • use simulation method: generate samples from the posterior and use them to estimate the quantities of interest
interval estimate	Confidence Interval: a claim that the region covers the true parameter, reflecting uncertainty in sampling procedure. e.g: 95% CI=(a,b) implies the interval (a,b) covers the true parameter among 95% of the experiments.	Credible Interval: a claim that the true parameter is inside the region with measurable probability. One can make a direct probability statement about parameters. e.g: 95% CI=(a,b) implies the chance that the true parameter falls in (a,b) is 95%.

1.5.1 Advantages Of Bayesian Inference Over Frequentist Inference

Following is a short list of advantages of Bayesian inference over frequentist inference.

- Bayesian inference allows informative priors so that prior knowledge or results of a previous model can be used to inform the current model.
- Bayesian inference considers the data to be fixed (which it is), and parameters to be random because they are unknowns. Frequentist inference considers the unknown parameters to be fixed, and the data to be random, estimating not based on the data at hand, but the data at hand plus hypothetical repeated sampling in the future with similar data. Bayesian inference provides answers conditional on the observed data and not based on the distribution of estimators or test statistics over imaginary samples not observed.
- Bayesian inference includes uncertainty in the probability model, yielding more realistic predictions. Frequentist inference does not include uncertainty of the parameter estimates, yielding less realistic predictions.
- Bayesian inference estimates a full probability model. Frequentist inference does not. There is no frequentist probability distribution associated with parameters or hypotheses.
- Bayesian inference estimates P(hypothesis|data). In contrast, frequentist inference estimates P(data|hypothesis). Even the term 'hypothesis testing' suggests it should be the hypothesis that is tested, given the data, not the other way around.
- Bayesian inference has an axiomatic foundation that is uncontested by frequentists. Therefore, Bayesian inference is coherent to a frequentist, but frequentist inference is incoherent to a Bayesian.
- Bayesian inference has a decision theoretic foundation. The purpose of most of statistical inference is to facilitate decision making. The optimal decision is the Bayesian decision.
- Bayesian inference includes uncertainty in the probability model, yielding more realistic predictions. Frequentist inference does not include uncertainty of the parameter estimates, yielding less realistic predictions.

- Bayesian inference uses observed data only. Frequentist inference uses both observed data and future data that is unobserved and hypothetical.
- Bayesian inference via MCMC or algorithms allows more complicated models that frequentists are unable to estimate.
- Bayesian inference via MCMC is unbiased with respect to sample size and can accommodate any sample size no matter how small. Frequentist inference becomes more biased as sample size decreases from infinity, and is often wildly biased with small samples, so minimum sample size is an issue.

See [16].

Chapter 2

Finite Mixture of Distributions

2.1 Introduction

In this chapter we will give an introduction to finite mixture models. Also, we will give some applications of finite mixture distributions. Then, we set up the finite mixture models and define the mixture density. After that, we introduce quick review for the Poisson distribution and its important properties. Finally, we present the finite Poisson mixture models using the missing data formulation.

2.2 Finite Mixture Models and Some Applications

Mixed distributions are widely used to model data in which each observation is assumed to come from one of a number of distributions with different parameters. In other words, we can say that mixture models are used when the population of sampling consists of number of subpopulations which have different parameters. Moreover, mixture models arise in practical problems when the measurements of a random variable are taken under two different conditions, for example, the distribution of heights in a population of adults reflects the mixture of males and females in the population, here the best way is to model male and female heights as separate univariate perhaps normal distributions rather than a single binomial distributions.

Note 2.2.1. This kind of mixture models are used when observations can be obtained only from the whole population and not from the components of the different populations.

2.2.1 Some Applications of Finite Mixtures

Finite mixture distributions arise in a variety of applications ranging from the length distribution of fish to the content of DNA in the nuclei of liver cells.

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.

Gregor (1969) applies a mixture of normal distributions to data arising from measuring the content of DNA in the nuclei of liver cells of rats. Such a distribution is through appropriate in this case because in some organs there exist various classes of nuclei of cells which have characteristic differences in DNA content.

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.
He use mixture distributions to know if there is more than one type of a disease.
Clark et al. studying hypertension, investigate whether a sample of blood pressure data can be separated into two normal populations.

Another general area where mixtures of distributions are important is in failure data. Here the observations are the times of failure of a sample of items. Often failure can occur for more than one reason, and the failure distribution for each reason can be adequately approximated by a simple density function such as the negative exponential. The overall failure distribution is then a mixture. Several attempts have been made to fit such mixtures to the failure distribution of electronic values.

Discrete mixtures are applied by Medgyessi (1961) to the counter current method of identifying the constituents of organic chemicals. This involves sequentially diffusing the dissolved chemical mixture

into a number of cells containing fresh solving. The result of this exercise is that each component of the chemical mixture is distributed independently of others according to binomial distributions across the cells, the final result is binomial mixture. See [7].

2.3 Setting Up Mixture Models

In Section 2.2 we gave an informal description of the mixture models and their applications. In this section we begin with a more formal definition of a mixture density.
We start by comparing two different graphical models:

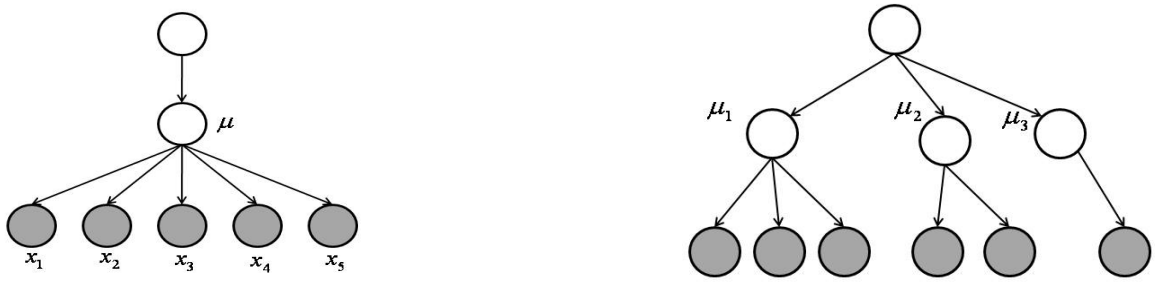

Figure 2.1: Graphical models

Firstly in the left side of the above figure we have a normal case at which we drown a sample from a population with a parameter (may be more than one) represents the whole population. In this case the estimation possible by standard methods such as a maximum likelihood method.

On the other side of the figure a sample drown from a nonhomogeneous population which divided to three subpopulations (or clusters) each one has a different parameter form the other. Here we do not know each observation x_i belong to any one of these clusters. We call this case a mixture model with three clusters.
For a mixture model, estimation of the parameters and the cluster structure require more advanced methods such as Bayesian Inference and this is what we are introducing it later.

Definition 2.3.1. [8] A random variable X is said to have a finite mixture distribution if its density function f can be expressed in the form

$$f(x|\theta, p) = \sum_{j=1}^{k} p_j f(x|\theta_j) \tag{2.1}$$

where $\theta = (\theta_1, \theta_2, ..., \theta_k)$, $p = (p_1, p_2, ..., p_k)$ for some probabilities $p_j > 0$, $j = 1, ..., k$, $k > 1$, with

$$\sum_{j=1}^{k} p_j = 1.$$

The above random variable is said to have a $k-$finite mixture density.

Remark 2.3.1. In equation 2.1 above,

1. θ_j is either a vector of parameters or a scaler referring to the j^{th} component of the mixture, and θ_j is called the mixing parameter or the parameter of the j^{th} component.

2. Usually, the p_j's are called the mixing proportions or weights and are most often unknown, and one p_j equal the proportion of that x belonging to the j^{th} component.

3. The number of components k may be unknown, so it is considered as a random variable, our interest is of finite or known number of components only.

2.4 Finite Poisson Mixture

Mixed poisson distributions are widely used in various disciplines to model data in which each observation is assumed to come from one of a number of poisson distributions with different parameters. In this section we will present the finite poisson mixture model using the missing data formulation.

2.4.1 The Poisson Distribution

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

Definition 2.4.1. [20] Let X be the number of events in a given interval with λ mean number of events per interval. Then the probability of observing x events in a given interval is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, 3, 4, ... \tag{2.2}$$

If the probabilities of X are distributed in this way then we say that X has a Poisson Distribution with parameter λ and we write $X \sim Poisson(\lambda)$.

Note that for a $Poisson(\lambda)$, mean = variance $=\lambda$.

2.4.2 Set up Finite Poisson Mixtures

Definition 2.4.2. [5] The probability function of the $k-$finite Poisson mixture is given by

$$f(x|\lambda, p) = \sum_{j=1}^{k} p_j \frac{e^{-\lambda_j}\lambda_j^x}{x!} \tag{2.3}$$

where $p = (p_1, p_2, ..., p_k)$, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ for some probabilities $p_j > 0$, $j = 1, ..., k$, $k > 1$, with

$$\sum_{j=1}^{k} p_j = 1.$$

We assume that $\lambda_1 < \lambda_2 < ... < \lambda_k$ to ensure the identifiability of the above finite mixture.

Remark 2.4.1. [32] 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 $\theta_j = \theta'_{\pi_j}$.

2.4.3 Missing Data

It is convenient to introduce the missing data formulation of the model, in which each observation x_i is assumed to arise from a specific but unknown (that is, missing) component of the mixture.
This formulation arises in the following context.

Let $X_1, X_2, ..., X_n$ be iid distributed according to Equation 2.3, with the parameters vector λ $=(\lambda_1, ..., \lambda_k)$ and proportions $p = (p_1, ..., p_k)$.

In this situation, for each observation $x_i, i = 1, 2, ..., n$, the indicator parameter z_i is introduced such as $z_{ij} = 1$ indicates that observation x_i belongs to the j^{th} component of the mixture.
That is

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

Thus the density $f(x_i|z_{ij} = 1)$ is Poisson (λ_j) and $f(z_{ij} = 1|p) = p_j$.
Given $p = (p_1, ..., p_k)$, the distribution of each unobserved vector $z_i = (z_{i1}, ..., z_{ik})$ is Multinomial$(1, p_1, ..., p_k)$.
For more details see chapter 5. See [8].

Now the question is, why we use the missing data approach?

To see this, consider the case of n iid observations $x = (x_1, ..., x_n)$ from the mixture model (see Equation 2.1) . Defining $p = (p_1, ..., p_k)$ and $\theta = (\theta_1, ..., \theta_k)$, The likelihood density of the mixture is:

$$\begin{aligned} f(x|\theta, p) &= \prod_{i=1}^{n} f(x_i|\theta, p) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i|\theta_j) \end{aligned}$$

the full computation of the posterior distribution and in particular the explicit representation of the corresponding posterior expectation involves the expansion of the likelihood of the mixture into k^n terms, which is computationally too expensive to be used for more than a few observations.
So the missing data representation of a mixture distribution can be exploited as a technical device to facilitate numerical estimation.

Now if we look at the expansion of the above likelihood we will see that each observation x_i adds a single factor to the product namely when $j = i$.

So we can use the missing indicator z_i with exactly one of z_{ij} equaling 1 for each i , and we can write the likelihood of the mixture as

$$f(x, z|\theta, p) = \prod_{i=1}^{n} \sum_{j=1}^{k} z_{ij} p_j f(x_i|\theta_j) \tag{2.4}$$

we call this the complete data likelihood. See [1].
Equation 2.4 can be written as an equivalence formula which not involves summation that is

$$f(x, z|\theta, p) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j f(x_i|\theta_j))^{z_{ij}}. \quad \text{See[8].} \tag{2.5}$$

Where $f(x_i|\theta_j)$ is in our case Poisson distribution with parameter λ_j.

To understand how we get the equation 2.5 see the following example.

Example 2.4.1. 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 a Poisson distribution.

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

Table 2.1: The data points and their clusters.

	x_i	k
1	x_1	1
2	x_2	2
3	x_3	1
4	x_4	2

The complete data likelihood equals

$$
\begin{aligned}
f(x, z|\theta, p) &= \prod_{i=1}^{n}\sum_{j=1}^{k} z_{ij}p_j f(x_i|\theta_j) \\
&= [z_{11}f(x_1|\lambda_1)p_1 + z_{12}f(x_1|\lambda_2)p_2].[z_{21}f(x_2|\lambda_1)p_1 + z_{22}f(x_2|\lambda_2)p_2]. \\
&[z_{31}f(x_3|\lambda_1)p_1 + z_{32}f(x_3|\lambda_2)p_2].[z_{41}f(x_4|\lambda_1)p_1 + z_{42}f(x_4|\lambda_2)p_2] \\
&= [z_{11}f(x_1|\lambda_1)p_1 + 0].[0 + z_{22}f(x_2|\lambda_2)p_2].[z_{31}f(x_3|\lambda_1)p_1 + 0)p_2].[0 + z_{42}f(x_4|\lambda_2)p_2] \\
&= \frac{e^{-\lambda_1}\lambda_1^{x_1}}{x_1!}p_1\frac{e^{-\lambda_2}\lambda_2^{x_2}}{x_2!}p_2\frac{e^{-\lambda_1}\lambda_1^{x_3}}{x_3!}p_1\frac{e^{-\lambda_2}\lambda_2^{x_4}}{x_4!}p_2 \qquad (z_{11} = z_{22} = z_{31} = z_{42} = 1) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k} z_{ij}p_j f(x_i|\theta_j) \\
&\equiv \prod_{i=1}^{n}\prod_{j=1}^{k}(p_j f(x_i|\theta_j))^{z_{ij}}
\end{aligned}
$$

Note that the remaning terms in the summation for each observation is only one term that is when $i = j$ so that we could replace summation by product.

Chapter 3

Markov Chains

Markov chains were introduced in 1906 by Andrei Andreyevich Markov (1856 - 1922) and were named in his honor. Markov is particularly remembered for his study of Markov chains. His research works on Markov chains launched the study of stochastic processes with a lot of applications.

Markov chains are the simplest mathematical models for random phenomena evolving in time. Their simple structure makes it possible to say a great deal about their behaviour. At the same time, the class of Markov chains is rich enough to serve in many applications, they are widely used in various scientific areas such as finance and insurance or even in physics, chemistry or biology where one might wouldn't expect it at the first place. This makes Markov chains the first and most important examples of random processes.

A Markov process is a random process for which the future (the next step) depends only on the present state; it has no memory of how the present state was reached, this specific kind of "memorylessness" is called the **Markov property**.

In this chapter we will discuss some basic properties of a Markov chain. Basic concepts and notations are explained also some important theorems in this area will be presented.

3.1 Stochastic Processes

Markov chains are a general class of stochastic models. We first define stochastic processes generally, and then show how one finds discrete time Markov chains as probably the most intuitively simple class of stochastic processes.

Definition 3.1.1. [35] A stochastic process is a collection of random variables (on some probability space) indexed by some set I : $(X_n, n \in I)$. When $I \subseteq \mathbf{R}$ we can think of I as a set of points in time, and X_n as the state of the process at time n. The state space, denoted by $\mathcal{S}$, is the set of all possible values of the X_n. When I is countable we have a discrete-time stochastic process. When I is an interval of the real line we have a continuous-time stochastic process.

Example 3.1.1. [9] Let X_n be the number of customers served in a bank at the end of the nth working day. Then $\{X_n : n = 1, 2, \ldots\}$ is a stochastic process. It is called a discrete-time stochastic process since its index set, $I = \{1, 2, \ldots\}$, is countable. The state space for the number of customers served in a bank at the end of the nth working day is $\mathcal{S} = \{0, 1, 2, \ldots\}$.

Example 3.1.2. [9] Suppose that there are three machines in a factory, each working for a random time that is exponentially distributed. When a machine fails, the repair time is also a random variable exponentially distributed. Let X(t) be the number of functioning machines in the factory at time t. Then $\{X(t) : t \geq 0\}$ is a continuous-time stochastic process with state space $\mathcal{S} = \{0, 1, 2, 3\}$.

3.2 Basic definitions and properties

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 next figure. See [11].

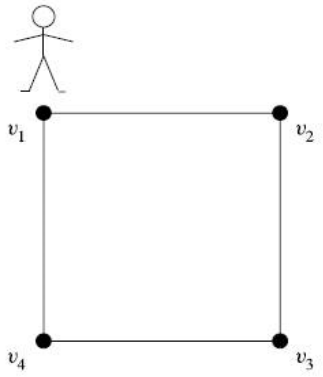

Figure 3.1: A random walker in a very small town.

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 the above figure, 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, $(X_0, X_1, \ldots)$ 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 $\frac{1}{2}$ for each, so that

$$P(X_1 = 2) = \frac{1}{2}$$

and

$$P(X_1 = 4) = \frac{1}{2}.$$

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

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

and

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

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

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

and

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

for any choice of $i_0, \ldots, i_{n-1}$.

(This is because the coin flip at time $n+1$ is independent of all previous coin flips, and hence also independent of $X_0, \ldots, X_n$.)

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. Or in other words: to make the best possible prediction of what happens tomorrow (time $n+1$), we only need to consider what happens today (time n), as the past (times $0, \ldots, n-1$) gives no additional useful information.

Here $\{X_n : n = 0, 1, \ldots\}$ is called a Markov chain.

Next we will give a more formal definition for Markov chains.

Definition 3.2.1. [9] A stochastic process $\{X_n : n = 0, 1, \ldots\}$ with a finite or countably infinite state space $\mathcal{S}$ is said to be a Markov chain, if for all $i, j, i_0, \ldots, i_{n-1} \in \mathcal{S}$, and $n = 0, 1, 2, \ldots$,

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i). \tag{3.2}$$

The elements of the state space $\mathcal{S}$ are not necessarily nonnegative integers (or numbers). However, for simplicity, it is a common practice to label the elements of $\mathcal{S}$ by nonnegative integers. If $\mathcal{S}$ is finite, the Markov chain is called a finite Markov chain or a finite-state Markov chain. If $\mathcal{S}$ is infinite, it is called an infinite Markov chain or an infinite-state Markov chain.

The main property of a Markov chain, expressed by Equation 3.2, is called the Markovian property

of the Markov chain. Thus, by the **Markovian property**,
Given the state of the Markov chain at present (X_n), its future state (X_{n+1}) is independent of the past states $(X_{n-1}, \ldots, X_1, X_0)$.

Remark 3.2.1. The above definition is of a discrete-time Markov chain, there is also a continuous-time Markov chain which is not of our interests in this thesis. From now on Markov chain will mean discrete one.

Definition 3.2.2. [4] The transition probability p_{ij} is the conditional probability represents the probability that, the process will make a transition to state j given that currently the process is state i in one step. That is,

$$p_{ij} = P(X_{n+1} = j | X_n = i)$$

With, $p_{ij} \geq 0$ for all $i, j \in \mathcal{S}$.

A Markov chain with state space $\mathcal{S}$ is said to have **stationary transition probabilities**, if, for all $i, j \in \mathcal{S}$, p_{ij} does not depend on the time that the transition will occur. That is, if $P(X_{n+1} = j | X_n = i)$ is independent of n.

Definition 3.2.3. [4] The matrix containing the transition probabilities, p_{ij},

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & & & \end{pmatrix}$$

sometimes simply denoted by (p_{ij}), is called the one-step transition probability matrix of the Markov chain $\{X_n : n = 0, 1, \ldots\}$.

Notations 3.2.1. (about the above definition)

- The j^{th} entry in the i^{th} row is the probability of a transition from state $i - 1$ to state $j - 1$.
- For $i \in \mathcal{S}$, note that the probability of a transition from state i is $\sum_{j=0}^{\infty} p_{ij}$. Hence we must have $\sum_{j=0}^{\infty} p_{ij} = 1$; that is, the sum of the elements of each row of the transition probability matrix is 1.
- The sum of the elements of a column is not necessarily 1.

- In matrix theory, if all of the entries of a matrix are nonnegative and the sum of the entries of each row is 1, then it is called a Markov matrix.

Definition 3.2.4. [11] A transition graph is a useful way to picture a Markov chain. The transition graph consists of nodes representing the states of the Markov chain, and arrows between the nodes, representing transition probabilities.

For instance, the random walk example above is a Markov chain, with state space $\{1,\ldots,4\}$ and transition matrix,

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

with a transition graph,

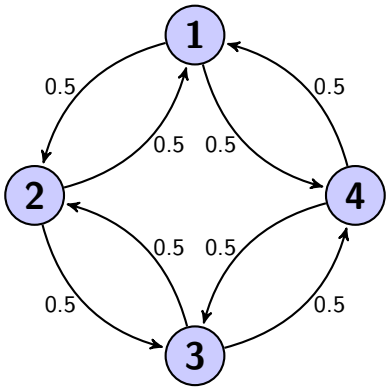

Figure 3.2: Transition graph of the random walk example.

Example 3.2.1. [11]**The Gothenburg weather.**
It is sometimes claimed that the best way to predict tomorrows weather is simply to guess that it will be the same tomorrow as it is today. If we assume that this claim is correct, then it is natural to model the weather as a Markov chain. For simplicity, we assume that there are only two kinds of weather: rain and sunshine. If the above predictor is correct 75% of the time (regardless of whether todays weather is rain or sunshine), then the weather forms a Markov chain with state space $\mathcal{S} = \{s_1, s_2\}$

(with s_1 = rain and s_2 = sunshine), and transition matrix,

$$\mathbf{P} = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$

With a transition graph,

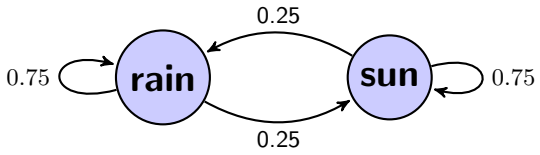

Figure 3.3: Transition graph of the Gothenburg weather example.

3.2.1 The initial distribution

We next consider another important characteristic (besides the transition matrix) of a Markov chain $(X_0, X_1, \ldots)$ with state space $\mathcal{S} = s_1, \ldots, s_k$, namely the initial distribution , which tells us how the Markov chain starts. **The initial distribution** is represented as a row vector $\mu^{(0)}$ given by

$$\begin{aligned} \mu^{(0)} &= (\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}) \\ &= (P(X_0 = s_1), P(X_0 = s_2), \ldots, P(X_0 = s_k)). \end{aligned}$$

Since $\mu^{(0)}$ represents a probability distribution, we have

$$\sum_{i=1}^{k} \mu_i^{(0)} = 1, \quad i = 1, 2, \ldots, k. \quad \text{See[11].}$$

In the random walk example above, we have

$$\mu^{(0)} = (1, 0, 0, 0)$$

because of Equation 3.1. Similarly, we let the row vectors $\mu^{(1)}, \mu^{(2)}, \ldots$ denote the distributions of the Markov chain at times 1, 2, ..., so that

$$\begin{aligned} \mu^{(n)} &= (\mu_1^{(n)}, \mu_2^{(n)}, \ldots, \mu_k^{(n)}) \\ &= (P(X_n = s_1), P(X_n = s_2), \ldots, P(X_n = s_k)). \end{aligned}$$

For the random walk example,

$$\mu^{(n)} = (0, \frac{1}{2}, 0, \frac{1}{2}).$$

3.3 The n-Step Transition Matrix

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

Definition 3.3.1. [4] Define p_{ij}^n to be the probability that a process in state i will be in state j after n additional transitions that is,

$$p_{ij}^n = P(X_{n+m} = j | X_m = i), \quad n, m \geq 0.$$

In particular $p_{ij}^1 = p_{ij}$.

Definition 3.3.2. [9] The matrix containing the n-step transition probabilities, p_{ij}^n,

$$\mathbf{P^{(n)}} = \begin{pmatrix} p_{00}^n & p_{01}^n & p_{02}^n & \cdots \\ p_{10}^n & p_{11}^n & p_{12}^n & \cdots \\ p_{20}^n & p_{21}^n & p_{22}^n & \cdots \\ \vdots & & & \end{pmatrix}$$

is called the ***n*-step transition probability matrix**.
Clearly, $\mathbf{P}^{(0)}$ is the identity matrix. That is, $p_{ij}^0 = 1$ if $i = j$, and $p_{ij}^0 = 0$ if $i \neq j$. Also, $\mathbf{P}^{(1)} = \mathbf{P}$, the transition probability matrix of the Markov chain.

3.3.1 Transition in $n + m$ Steps

For a transition from state i to state j in $n + m$ steps, we have to go from i to some state k in m steps and then from k to j in n steps, exactly as the next figure. The Markov property implies that the two parts of our journey are independent.
The above observation leads us to the following equations called the **Chapman-Kolmogorov** equations [6]:

$$p_{ij}^{m+n} = \sum_k p_{ik}^m \cdot p_{kj}^n. \tag{3.3}$$

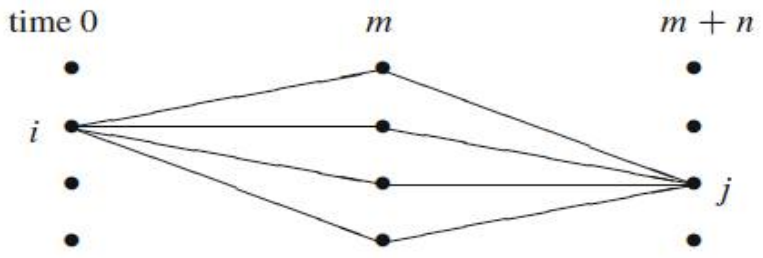

Figure 3.4: Transition in $m + n$ steps

Equation 3.3 can be proved by applying the definition of conditional probability,

$$\begin{aligned} p_{ij}^{m+n} &= P(X_{m+n} = j | X_0 = i) \\ &= \sum_k P(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_k \frac{P(X_{m+n} = j, X_m = k, X_0 = i)}{P(X_0 = i)} \\ &= \sum_k \frac{P(X_{m+n} = j, X_m = k, X_0 = i)}{P(X_m = k, X_0 = i)} \frac{P(X_m = k, X_0 = i)}{P(X_0 = i)} \\ &= \sum_k P(X_{m+n} = j | X_m = k, X_0 = i) P(X_m = k | X_0 = i) \end{aligned}$$

By Markov property the last expression is

$$\begin{aligned} &= \sum_k P(X_{m+n} = j | X_m = k) P(X_m = k | X_0 = i) \\ &= \sum_k p_{ik}^m \cdot p_{kj}^n. \end{aligned}$$

Note that in Equation 3.3,
p_{ij}^{n+m} is the ij^{th} entry of the matrix $\mathbf{P}^{(n+m)}$, p_{ik}^n is the ik^{th} entry of the matrix $\mathbf{P}^{(n)}$, and p_{kj}^m is the kj^{th} entry of the matrix $\mathbf{P}^{(m)}$. But, from the definition of the product of two matrices, the defining relation for the ij^{th} entry of the product of matrices $\mathbf{P}^{(n)}$ and $\mathbf{P}^{(m)}$ is identical to Equation 3.3.
Hence the **Chapman-Kolmogorov** equations, in matrix form, are

$$\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \cdot \mathbf{P}^{(m)} \tag{3.4}$$

which implies that,

$$\mathbf{P}^{(2)} = \mathbf{P}^{(1)} \cdot \mathbf{P}^{(1)} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2,$$
$$\mathbf{P}^{(3)} = \mathbf{P}^{(2)} \cdot \mathbf{P}^{(1)} = \mathbf{P}^2 \cdot \mathbf{P} = \mathbf{P}^3,$$

and, in general, by induction,

$$\mathbf{P}^{(n)} = \mathbf{P}^{(n-1)} \cdot \mathbf{P}^{(1)} = \mathbf{P}^{n-1} \cdot \mathbf{P} = \mathbf{P}^n. \quad \text{See[9].}$$

We have shown a very useful way to compute the n-step transition probability matrix:
The n-step transition probability matrix is equal to the one-step transition probability matrix raised to the power of n.

Example 3.3.1. [9] At 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 day n the traffic light will work; $X_n = 0$ if on day n it will not work. Then $\{X_n : n = 0, 1, \ldots\}$ is a Markov chain with state space {0, 1}.

$$p_{ij} = P(X_{n+1} = j | X_n = i),$$
$$p_{00} = P(X_{n+1} = 0 | X_n = 0) = 0.12$$
$$p_{01} = P(X_{n+1} = 1 | X_n = 0) = 0.88$$
$$p_{10} = P(X_{n+1} = 0 | X_n = 1) = 0.07$$
$$p_{11} = P(X_{n+1} = 1 | X_n = 1) = 0.93$$

so the transition probability matrix is,

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

the two-step transition probability matrix is given by

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

This shows that, for example, an out-of-order traffic light will be working the day after tomorrow with probability 0.924. Similarly, a working traffic light will be out of order the day after tomorrow with

probability 0.0735.

The matrix

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

shows that, whether or not the traffic light is working today, in six days, the probability that it will be working is 0.926316, and the probability that it will be out of order is 0.0736842.

Example 3.3.2. [9] In the model for the Gamblers ruin, suppose that player A's initial fortune is \$3 and player B's initial fortune is \$1. Furthermore, suppose that player A wins \$1 from B with probability 0.6 and loses \$1 to B with probability 0.4. Let X_n be player A's fortune after n games. Then the transition probability matrix of the Markov chain $\{X_n : n = 1, 2, \ldots\}$ is

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 & 0 \\ 0 & 0.4 & 0 & 0.6 & 0 \\ 0 & 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Direct calculations show that

$$\mathbf{P}^{(10)} = \mathbf{P}^{10} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.575 & 0.013 & 0 & 0.019 & 0.393 \\ 0.3 & 0 & 0.025 & 0 & 0.675 \\ 0.117 & 0.0085 & 0 & 0.0127 & 0.862 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Therefore, given that gambler A's initial fortune is \$3, after 10 games, the probability that A has, say, \$2 is 0; the probability that his fortune is \$3 is 0.0127; the probability that he wins the game (his fortune is \$4) is 0.862; and the probability that he loses the game (his fortune is \$0) is 0.117.

Let $\{X_n : n = 0, 1, \ldots\}$ be a Markov chain with its transition probability matrix given. The following theorem shows that if the probability mass function of X_0 is known, then, for all $n \geq 1$, we can find the probability mass function of X_n.

Theorem 3.3.1. *[9] Let $\{X_n : n = 0, 1, \ldots\}$ be a Markov chain with transition probability matrix $\mathbf{P} = (p_{ij})$. 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

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

Proof. Applying the law of total probability, Theorem 1.2.1 to the sequence of mutually exclusive events $\{X_0 = i\}, i \geq 0$, we have

$$\begin{aligned} P(X_n = j) &= \sum_{i=1}^{\infty} P(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i=1}^{\infty} p_{ij}^n p(i) \\ &= \sum_{i=1}^{\infty} p(i) p_{ij}^n. \end{aligned}$$

□

Example 3.3.3. Suppose that, in the Example of a random walker moves in a very small town consisting of four streets, and four street-corners v_1, v_2, v_3 and v_4, initially, it is equally likely that the walker is in any of the four street-corners. That is,

$$p(i) = P(X_0 = v_i) = \frac{1}{4}, \quad 1 \leq i \leq 4.$$

Then, using the matrix $\mathbf{P}^5$, which is the same as $\mathbf{P}$,

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

we can readily find the probability that the walker is in corner $i, 1 \leq i \leq 4$, after 5 transitions. For example,

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

3.4 Irreducible and Aperiodic Markov Chains

For several of the most interesting results in Markov theory, we need to put certain assumptions on the Markov chains we are considering. It is an important task, in Markov theory just as in all other branches of mathematics, to find conditions that on the one hand are strong enough to have useful consequences, but on the other hand are weak enough to hold (and be easy to check) for many interesting examples. In this section, we will discuss two such conditions on Markov chains: irreducibility and aperiodicity. These conditions are of central importance in Markov theory, and in particular they play a key role in the study of stationary distributions, which is the topic of the next section. [11].

Definition 3.4.1. [9] Let $\{X_n : n = 0, 1, \ldots\}$ be a Markov chain with state space $\mathcal{S}$ and transition probability matrix $\mathbf{P}$. 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,

$$i \to j \text{ if for some } n \geq 0,\ p_{ij}^n > 0.$$

If $p_{ij}^n = 0, \forall n \geq 0$, then we say j is **not accessible** from state i.

Definition 3.4.2. [9] If two states 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 $i \in \mathcal{S}, i \leftrightarrow i$ 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: If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$. To show this, we will establish that $i \to k$. The proof that $k \to i$ is similar. 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,

$$p_{ik}^{n+m} = \sum_{\ell=0}^{\infty} p_{i\ell}^n p_{\ell k}^m \geq p_{ij}^n p_{jk}^m > 0,$$

showing that $i \to k$.

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.
This takes us directly to the definition of irreducibility.

Definition 3.4.3. [11] A Markov chain $(X_0, X_1, \ldots)$ with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$ is said to be **irreducible** if for all $i, j \in \mathcal{S}$ we have that $i \leftrightarrow j$.
Or in other words a Markov chain is said to be **irreducible** if it has only one class. Otherwise the chain is said to be **reducible**.

Another way of phrasing the definition would be to say that the chain is irreducible if for any $i, j \in \mathcal{S}$ 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. Let us return to the transition graphs of the random walk example, and the Gothenburg weather example.

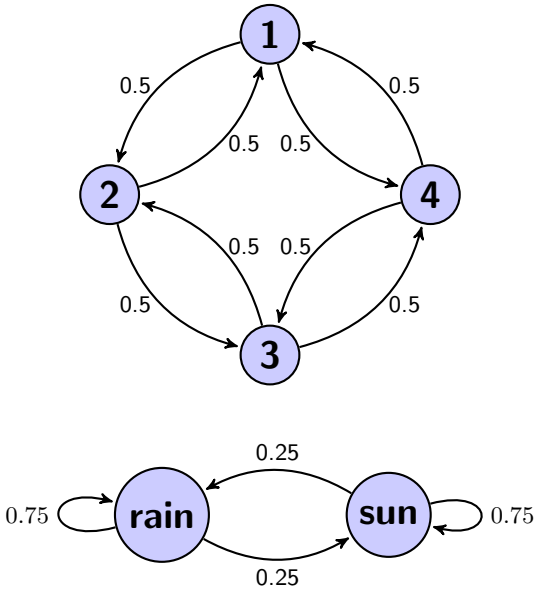

Figure 3.5: Transition graphs of irreducible Markov chains.

For the above transition graphs, we can easily check that from each state there is a sequence of arrows leading to any other state. So Markov chains of these examples are irreducible. Let us next have a look at an example which is not irreducible.

Example 3.4.1. [11] A reducible Markov chain.
Consider a Markov chain $(X_0, X_1, \ldots)$ with state space $\mathcal{S} = \{1, 2, 3, 4\}$ and transition matrix

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

By taking a look at its transition graph,

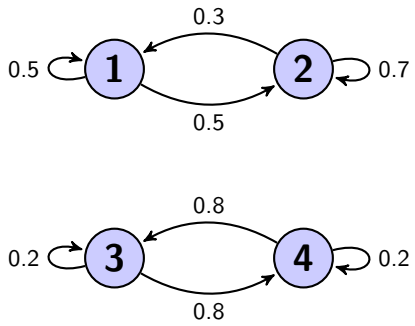

Figure 3.6: Transition graph of a reducible Markov chain.

we immediately see that if the chain starts in state 1 or state 2, then it is restricted to states 1 and 2 forever. Similarly, if it starts in state 3 or state 4, then it can never leave the subset {3, 4} of the state space. Hence, the chain is reducible. Note that if the chain starts in state 1 or state 2, then it behaves exactly as if it were a Markov chain with state space {1, 2} and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \end{pmatrix}$$

If it starts in state 3 or state 4, then it behaves like a Markov chain with state space {3, 4} and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}$$

This illustrates a characteristic feature of reducible Markov chains, which also explains the term reducible: If a Markov chain is reducible, then the analysis of it can be reduced to the analysis of one or more Markov chains with smaller state spaces.

We move on to consider the concept of **aperiodicity.**

For a finite or infinite set $\{a_1, a_2, \ldots\}$ of positive integers, we write $\gcd\{a_1, a_2, \ldots\}$ for the greatest common divisor of $a_1, a_2, \ldots$.

The period $d(i)$ of a state $i \in \mathcal{S}$ 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 (*i.e.*, has positive probability of returning) to i, given that we start with $X_0 = i$.

Definition 3.4.4. [11] Let $\{X_n : n = 0, 1, \ldots\}$ be a Markov chain with state space $\mathcal{S}$, for $i \in \mathcal{S}$, if the period $d(i) = \gcd\{n \geq 1 : p_{ii}^n > 0\} = 1$, then we say that the state i is **aperiodic.**

Definition 3.4.5. [11] A Markov chain is said to be **aperiodic** if all its states are aperiodic. Otherwise the chain is said to be **periodic**.

Consider for instance Example 3.2.1 **(the Gothenburg weather)**. It is easy to check that regardless of whether the weather today is rain or sunshine, we have for any n that the probability of having the same weather n days later is strictly positive. Or, expressed more compactly: $p_{ii}^n > 0$ for all n and all states i. This obviously implies that the Markov chain in Example 3.2.1 is **aperiodic.**

On the other hand, let us consider the random walk example, where the random walker stands in corner v_1 at time 0. Clearly, he has to take an even number of steps in order to get back to v_1. This means that $p_{11}^n > 0$ only for $n = 2, 4, 6, \ldots$. Hence, $\gcd\{n \geq 1 : p_{ii}^n > 0\} = \gcd\{2, 4, 6, \ldots\} = 2$, and the chain is therefore **periodic.**

Definition 3.4.6. [9] A state i of a Markov chain is called **absorbing** if $p_{ii} = 1$.

That is, once the chain visit (absorbed by) an absorbing state, it stays there forever and cannot move to any other state.

Definition 3.4.7. [24] A finite, irreducible, and aperiodic chain is called **regular** (or **ergodic**).

Returning to the example of the Gothenburg weather, we find its Markov chain irreducible, and aperiodic hence, it ergodic.

3.5 Recurrence and Transience

Definition 3.5.1. [9] For a Markov chain $\{X_n : n = 0, 1, \ldots\}$, define f_{ii}^n to 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. Clearly,

$$f_i = \sum_{n=1}^{\infty} f_{ii}^n.$$

Definition 3.5.2. [9] If $f_i = 1$, that is starting from i, the process returns to i with probability 1, then the state i is called **recurrent**.

Definition 3.5.3. [9] If $f_i < 1$ that is, if, starting from i, there is a positive probability that the process does not return to i, then the state i is called **transient**.

Remark 3.5.1. [9]

- Suppose that, starting from i, the process returns to i with probability 1. Since each time at i the process probabilistically restarts itself, the first return to i implies a second return to i, and so on. Therefore, if state i is recurrent, then the process enters i, with probability 1, infinitely many times.

- Starting from i, for a transient state i, the probability that the process return to i exactly n times is $(f_i)^n(1 - f_i), n \geq 0$.

  Thus the number of returns to i is a geometric random variable with parameter $1 - f_i$, and hence the average number of returns to i is $1/(1 - f_i) < \infty$. Therefore, if i is transient, then the average number of returns to i is finite.

Definition 3.5.4. [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 is called **null recurrent.**

The following theorem gives a tool to determine whether a state is transient or it is recurrent.

Theorem 3.5.1. *[4] For a Markov chain $\{X_n : n = 0, 1, \ldots\}$, with transition probability matrix $\mathbf{P} = (p_{ij})$,*

(a) State i is recurrent if and only if $\sum_{n=1}^{\infty} p_{ii}^n = \infty$.

(*b*) *State i is transient if and only if $\sum_{n=1}^{\infty} p_{ii}^n < \infty$.*

Proof. Firstly we want to show that, given $X_0 = i, \sum_{n=1}^{\infty} p_{ii}^n$ is the average number of returns to state i. To prove this, for $n \geq 1$, let

$$Z_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i. \end{cases}$$

The number of returns to state i is $\sum_{n=1}^{\infty} Z_n$, and we have

$$\begin{aligned} E\left(\sum_{n=1}^{\infty} Z_n | X_0 = i\right) &= \sum_{n=1}^{\infty} E(Z_n | X_0 = i) \\ &= \sum_{n=1}^{\infty} P(X_n = i | X_0 = i) \\ &= \sum_{n=1}^{\infty} p_{ii}^n. \end{aligned}$$

Now since the average number of returns to i is $\dfrac{1}{1 - f_i}$
so, by using the above result

$$\sum_{n=1}^{\infty} p_{ii}^n = \frac{1}{1 - f_i}$$

then $\sum_{n=1}^{\infty} p_{ii}^n = \infty$ iff $f_i = 1$ iff the state i is recurrent so we prove (a).
And $\sum_{n=1}^{\infty} p_{ii}^n < \infty$ iff $f_i < 1$ iff the state i is transient so (b) proved. □

Theorem 3.5.2. *[9] Recurrence is a class property. That is, if state i is recurrent and state j communicates with state i, then state j is also recurrent.*

Proof. Since i and j communicate, there exist n and m so that $p_{ij}^n > 0$ and $p_{ji}^m > 0$. Since i is recurrent, $\sum_{k=1}^{\infty} p_{ii}^k = \infty$. For $k \geq 1$, applying Chapman-Kolmogorov equations repeatedly yields

$$p_{jj}^{n+m+k} = \sum_{\ell=0}^{\infty} p_{j\ell}^m p_{\ell j}^{n+k} \geq p_{ji}^m p_{ij}^{n+k} = p_{ji}^m \sum_{\ell=0}^{\infty} p_{i\ell}^k p_{\ell j}^n \geq p_{ji}^m p_{ii}^k p_{ij}^n.$$

Hence

$$\sum_{k=1}^{\infty} p_{jj}^{n+m+k} \geq \sum_{k=1}^{\infty} p_{ji}^m p_{ii}^k p_{ij}^n = p_{ji}^m p_{ij}^n \sum_{k=1}^{\infty} p_{ii}^k = \infty,$$

since $p_{ji}^m > 0, p_{ij}^n > 0$, and $\sum_{k=1}^{\infty} p_{ii}^k = \infty$. This implies that $\sum_{k=1}^{\infty} p_{jj}^{n+m+k} = \infty$ which gives

$$\sum_{k=1}^{\infty} p_{jj}^k \geq \sum_{k=1}^{\infty} p_{jj}^{n+m+k} = \infty,$$

or $\sum_{k=1}^{\infty} p_{jj}^k = \infty$. Hence j is recurrent as well. □

Theorem 3.5.3. *[9] Transience is a class property. That is, if state i is transient, and state j communicates with state i, then state j is also transient.*

Proof. Suppose that j is not transient; then it is recurrent. Since j communicates with i, by previous theorem, i must also be recurrent; a contradiction. Therefore, state j is transient as well. □

As a result of the facts that transience and recurrence are class properties, we have that in an irreducible Markov chain, either all states are transient, or all states are recurrent.

3.6 Stationary Distribution

In this chapter, we consider one of the central issues in Markov theory which is called stationary distribution.

Definition 3.6.1. [11] Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\{i_1, \ldots, i_k\}$ and transition matrix $\mathbf{P}$. A row vector $\pi = (\pi_1, \ldots, \pi_k)$ is said to be a stationary distribution for the Markov chain, if it satisfies

(i) $\pi_i \geq 0$ for $i = 1, \ldots, k$,

(ii) $\sum_{i=1}^{k} \pi_i = 1$,

(iii) $\pi\mathbf{P} = \pi$ meaning that $\sum_{i=1}^{k} \pi_i p_{ij} = \pi_j$ for $j = 1, \ldots, k$.

Example 3.6.1. Consider a Marcov chain with a transition matrix,

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

The equation $\pi p = \pi$ says

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

which translates into three equations

$$0.7\pi_1 + 0.3\pi_2 + 0.2\pi_3 = \pi_1$$
$$0.2\pi_1 + 0.5\pi_2 + 0.4\pi_3 = \pi_2$$
$$0.1\pi_1 + 0.2\pi_2 + 0.4\pi_3 = \pi_3$$

Note that the columns of the matrix give the numbers in the rows of the equations. The third equation is redundant since if we add up the three equations we get

$$\pi_1 + \pi_2 + \pi_3 = \pi_1 + \pi_2 + \pi_3$$

If we replace the third equation by $\pi_1+\pi_2+\pi_3 = 1$ and subtract π_1 from each side of the first equation and π_2 from each side of the second equation we get

$$-0.3\pi_1 + 0.3\pi_2 + 0.2\pi_3 = 0$$
$$0.2\pi_1 - 0.5\pi_2 + 0.4\pi_3 = 0$$
$$\pi_1 + \pi_2 + \pi_3 = 1$$

At this point we can solve the equations by hand or using a calculator.

By hand. We note that the third equation implies $\pi_3 = 1 - \pi_1 - \pi_2$ and substituting this in the first two gives

$$0.2 = 0.5\pi_1 - 0.1\pi_2$$
$$0.4 = 0.2\pi_1 + 0.9\pi_2$$

Multiplying the first equation by 0.9 and adding 0.1 times the second gives

$$2.2 = (0.45 + 0.02)\pi_1 \quad \text{or} \quad \pi_1 = \tfrac{22}{47}$$

Multiplying the first equation by 0.2 and adding -0.5 times the second gives

$$-0.16 = (-0.02 - 0.45)\pi_2 \quad \text{or} \quad \pi_2 = \tfrac{16}{47}$$

Since the three probabilities add up to 1, $\pi_3 = \frac{9}{47}$. So the stationary distribution $\pi = (\frac{22}{47}, \frac{16}{47}, \frac{9}{47})$.

3.6.1 Doubly Stochastic Chains

Definition 3.6.2. [6] A transition matrix $\mathbf{P}$ is said to be **doubly stochastic** if its columns sum to 1, or in symbols $\sum_i p_{ij} = 1$.

The adjective doubly refers to the fact that by its definition a transition probability matrix has rows that sum to 1, i.e., $\sum_j p_{ij} = 1$.

Theorem 3.6.1. *[6] If* $\mathbf{P}$ *is a doubly stochastic transition probability for a Markov chain with* N *states, then the uniform distribution,* $\pi(i) = \frac{1}{N}$ *for all* i*, is a stationary distribution.*

Proof. To check this claim we note that if $\pi(i) = \frac{1}{N}$ then

$$\sum_i \pi(i)p_{ij} = \frac{1}{N}\sum_i p_{ij} = \frac{1}{N} = \pi(j).$$

Therefore $\pi(i) = \frac{1}{N}$ is a stationary distribution. □

Example 3.6.2. [9] For an English course, there are four popular textbooks dominating the market. The English department of an institution allows its faculty to teach only from these four textbooks. Each year, Professor Rosemary O'Donoghue adopts the same book she was using the previous year with probability 0.64. The probabilities of her changing to any of the other three books are equal. Find the proportion of years Professor O'Donoghue uses each book.
Solution: For $1 \leq i \leq 4$, let $X_n = i$ if on year n Professor O'Donoghue teaches from book i. Then $\{X_n : n = 0, 1, 2, \ldots\}$ is a Markov chain with state space {1, 2, 3, 4} and transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 0.64 & 0.12 & 0.12 & 0.12 \\ 0.12 & 0.64 & 0.12 & 0.12 \\ 0.12 & 0.12 & 0.64 & 0.12 \\ 0.12 & 0.12 & 0.12 & 0.64 \end{pmatrix}$$

Clearly, $\{X_n : n = 0, 1, 2, \ldots\}$ is doubly stochastic. That is, in addition to the sum of the entries of each row of $\mathbf{P}$ being 1, the sum of the entries of each column of $\mathbf{P}$ is 1 as well. Therefore, by the above theorem, for $1 \leq i \leq 4$, the proportion of years Professor O'Donoghue uses book i is 1/4.

3.7 Reversible Markov Chains and Detailed Balance Condition

In this section we introduce a special class of Markov chains known as the reversible ones. They are called so because they, in a certain sense, look the same regardless of whether time runs backwards or forwards. We jump right on to the definition.

Definition 3.7.1. [18] A Markov chain $(X_n, n = 0, 1, \ldots)$ is said to be **reversible** if it backwards, and turns out to be also a Markov chain. That is,

$$P(X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n) = P(X_0 = x_n, X_1 = x_{n-1}, \ldots, X_{n-1} = x_1, X_n = x_0)$$

Definition 3.7.2. [11] Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. A probability distribution π on $\mathcal{S}$ is said to satisfy the **detailed balance condition** or **(reversible)** if for all $i, j \in \mathcal{S}$ we have

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{3.5}$$

Theorem 3.7.1. *[18] Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\mathcal{S}$ and transition matrix $\boldsymbol{P}$. If π is a reversible distribution for the chain, then it is also a stationary distribution for the chain.*

Proof. Property $(i), (ii)$ of Definition 3.6.1 is immediate, so it only remains to show that for any $j \in \mathcal{S}$, we have

$$\pi_j = \sum_{i=1}^{k} \pi_i p_{ij}.$$

We get

$$\pi_j = \pi_j \sum_{i=1}^{k} p_{ji} = \sum_{i=1}^{k} \pi_j p_{ji} = \sum_{i=1}^{k} \pi_i p_{ij}$$

where the last equality uses that π satisfy the detailed balance condition. □

Many chains do not have stationary distributions that satisfy the detailed balance condition, see the next example.

Example 3.7.1. [6] Consider

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.3 & 0.1 & 0.6 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} \end{array}$$

$p_{13} = 0$, so for any stationary distribution π,

$$\pi(1)p_{13} = 0 \text{ but, } p_{31} > 0 \text{ so, we must have } \pi(3) = 0.$$

And using $\pi(3)p_{3i} = \pi(i)p_{i3}$ we conclude all the $\pi(i) = 0$ which contradicts that π is stationary probability distribution. So there is no stationary distribution with detailed balance. Also this chain is doubly stochastic so $(1/3, 1/3, 1/3)$ is a stationary distribution.

The rest of this section will deal with three issues: the existence of stationary distributions, the uniqueness of stationary distributions, and the convergence to stationarity starting from any initial distribution. We will give the following three theorems without proofs, these theorems are very important in the next chapter.

Theorem 3.7.2. *[11]* ***(Existence of stationary distributions)*** *For any irreducible and aperiodic (ergodic) Markov chain, there exists at least one stationary distribution.*

Proof. see [11]. □

Theorem 3.7.3. *[11]****(Uniqueness of the stationary distribution)*** *Any irreducible and aperiodic Markov chain has exactly one stationary distribution.*

Proof. see [11]. □

Theorem 3.7.4. *[11]* ***(The Markov chain convergence theorem)*** *Let $(X_0, X_1, \ldots)$ be an irreducible aperiodic Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$, transition matrix $\mathbf{P}$, and arbitrary initial distribution $\mu^{(0)}$. Then, for any distribution π which is stationary for the transition matrix $\mathbf{P}$, we have*

$$\mu^{(n)} \to \pi$$

Proof. see [11]. □

The last theorem says that if we run a Markov chain for a sufficiently long time n, then, regardless of what the initial distribution was, the distribution at time n will be close to the stationary distribution π. This is often referred to as the Markov chain approaching equilibrium as $n \to \infty$.

Since the resulting models of the Markov chains are often too difficult to be analyzed analytically, computers are used for inference. So in the next chapter we will introduce the concept of Markov chain Monte Carlo and algorithms for Markov chains simulation.

Chapter 4

Markov Chain Monte Carlo Methods

4.1 Introduction

In this chapter we will look at Markov chain Monte Carlo (MCMC) methods for generating samples from the posterior distribution. Here we don't draw our sample from the posterior distribution directly, but, we will set up a Markov chain that has the posterior distribution as its stationary distribution. The Metropolis-Hastings (M-H) algorithm and Gibbs sampler are methods of doing this. In this chapter we present the algorithm used to generate samples.

4.2 Monte Carlo Sampling From The Posterior

In Bayesian statistics, we have two sources of information, our prior $f(\theta)$ and the observed data $f(\mathbf{x}|\theta)$. And as we saw in Chapter 1 Bayes' theorem combines the two sources into a single distribution after we have observed the data. The final distribution is known as the posterior distribution. Bayes' Theorem is usually expressed very simply in the unscaled form posterior proportional to prior times likelihood. In equation form this is

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta). \tag{4.1}$$

This formula does not give the posterior density $f(\theta|\mathbf{x})$ exactly, but it does give its shape. In other words, we can find where the modes are, and relative values at any two locations. However, it does not give the scale factor needed to make it a density. This means we cannot calculate probabilities or moments from it. Thus it is not possible to do any inference about the parameter θ from the unsealed posterior. The actual posterior density is found by scaling it so it integrates to one.

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)f(\theta)d\theta} \tag{4.2}$$

A closed form for the integral in the denominator only exists for some particular cases. For other cases the posterior density has to be approximated numerically. This requires integrating

$$\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)f(\theta)d\theta$$

numerically, which may be very difficult, particularly when the parameter θ is high dimensional.
The computational approach to Bayesian statistics allows the posterior to be approached from a completely different direction. Bayesian approach does not use the computer to calculate the posterior numerically, but, use the computer to draw a Monte Carlo sample from the posterior. Fortunately, all we need to know is the shape of the posterior density, which is given by the prior times the likelihood. We do not need to know the scale factor necessary to make it the exact posterior density. These methods replace the very difficult numerical integration with the much easier process of drawing random samples. A Monte Carlo random sample from the posterior will approximate the true posterior when the sample size is large enough. We will base our inferences on the Monte Carlo random sample from the posterior, not from the numerically calculated posterior. Sometimes this approach to Bayesian inference is the only feasible method, particularly when the parameter space is high dimensional. [2].

4.3 Metropolis-Hastings Algorithm

The Metropolis Hastings algorithm can be seen as one of the most general Markov chain Monte Carlo (MCMC) algorithms, it is also one of the simplest both to understand and explain, making it an ideal algorithm to start with. It works by reshaping a random sample drawn from an easily sampled candidate distribution (sometimes called the starting distribution) into a random sample from the posterior by only accepting some of the candidate values into the final sample. [26].

Generally we will only know the unsealed posterior density $g(\theta|x) \propto g(\theta) \times f(x|\theta)$, not the exact posterior. Fortunately, we will see that the unsealed posterior is all we need to know to find a Markov chain that has the exact posterior as its stationary distribution. After the chain has run for a large number of steps, a draw from the chain can be considered to be a random draw from the posterior.

We will use the easily sampled **candidate density** $q(\theta, \theta')$. It is very important that the candidate density dominates the unsealed posterior. That means that we can find a number M such that

$$M \times q(\theta, \theta') \geq g(\theta) \times f(x|\theta)$$

There are two kinds of candidate densities we can use, random walk candidate densities, or independent candidate densities.

Let $q(\theta, \theta')$ be a candidate distribution that generates a candidate θ' given starting value θ. If for all θ, θ' the candidate distribution $q(\theta, \theta')$ satisfies the reversibility condition

$$g(\theta|x) \times q(\theta, \theta') = g(\theta'|x) \times q(\theta', \theta)$$

then the candidates $q(\theta', \theta)$ give a Markov chain with $g(\theta'|x)$ as it's stationary distribution. See Theorem 3.7.1. See [2].

4.3.1 Metropolis-Hastings Algorithm for A single Parameter

Unfortunately, most candidate distributions don't satisfy the reversibility condition.

For some θ and θ'

$$g(\theta|x) \times q(\theta, \theta') \neq g(\theta'|x) \times q(\theta', \theta)$$

Metropolis et al. supplied the solution. They restored the balance by introducing a probability of moving

$$\alpha(\theta, \theta') = min\left[1, \frac{g(\theta'|x)q(\theta', \theta)}{g(\theta|x)q(\theta, \theta')}\right]. \tag{4.3}$$

We do not need to know the exact posterior. If we multiply the posterior $g(\theta|x)$ by a constant k, the factor k occurs in both the numerator and the denominator so it cancels out. The algorithm only requires that we know the unsealed posterior.

Similarly, we can multiply the candidate density by a constant, and since it occurs in both numerator and denominator it will also cancel out.

All we need is the part that gives the shape of the candidate density. **Hastings** made some significant improvements and extended the algorithm so now it is known as the **Metropolis - Hastings algorithm.** [2].

Theorem 4.3.1. *[2]The revised candidate distribution $\alpha(\theta, \theta') \times q(\theta, \theta')$ satisfies the reversibility condition. Where $\alpha(\theta, \theta')$ is the acceptance probability*

$$\alpha(\theta, \theta') = min\left[1, \frac{g(\theta'|x)q(\theta', \theta)}{g(\theta|x)q(\theta, \theta')}\right].$$

Proof.

$$\begin{aligned} g(\theta|x)\alpha(\theta, \theta')q(\theta, \theta') &= g(\theta|x) \times min\left[1, \frac{g(\theta'|x)q(\theta', \theta)}{g(\theta|x)q(\theta, \theta')}\right] q(\theta, \theta') \\ &= min\left[g(\theta|x)q(\theta, \theta'), g(\theta'|x)q(\theta', \theta)\right] \end{aligned}$$

$$\begin{aligned} g(\theta'|x)\alpha(\theta', \theta)q(\theta', \theta) &= g(\theta'|x) \times min\left[1, \frac{g(\theta|x)q(\theta, \theta')}{g(\theta'|x)q(\theta', \theta)}\right] q(\theta', \theta) \\ &= min\left[g(\theta'|x)q(\theta', \theta), g(\theta|x)q(\theta, \theta)\right]. \end{aligned}$$

So we have the reversibility condition

$$g(\theta|x)\alpha(\theta, \theta')q(\theta, \theta') = g(\theta'|x)\alpha(\theta', \theta)q(\theta', \theta).$$

□

Steps of Metropolis-Hastings Algorithm [2]

1. Start at an initial value $\theta^{(0)}$.
2. Do from $n = 1, \ldots, N$.
   - **(a)** Draw θ' from $q(\theta^{(n-1)}, \theta')$.
   - **(b)** Calculate the probability $\alpha(\theta^{(n-l)}, \theta')$.

(c) Draw u from $U(0,1)$ where $U \sim$ uniform.

(d) if $u < \alpha(\theta^{(n-l)}, \theta')$ then let $\theta^{(n)} = \theta'$, else let $\theta^{(n)} = \theta^{(n-1)}$.

In fact, when the candidate density is exactly the same shape as the posterior

$$q(\theta, \theta') = k \times g(\theta'|x)$$

the acceptance probability

$$\begin{aligned}\alpha(\theta, \theta') &= min\left[1, \frac{g(\theta'|x)q(\theta', \theta)}{g(\theta|x)q(\theta, \theta')}\right] \\ &= min\left[1, \frac{g(\theta'|x)g(\theta|x)}{g(\theta|x)g(\theta'|x)}\right] \\ &= 1\end{aligned}$$

Thus, in that case, all candidates will be accepted.

Single Parameter with a Random-Walk Candidate Density

For a random-walk candidate generating distribution, the candidate is drawn from a symmetric distribution centered at the current value. Thus the candidate density is given by

$$q(\theta, \theta') = q_1(\theta' - \theta)$$

where $q_1()$ is a function symmetric about 0. Because of the symmetry $q_1(\theta' - \theta) = q_1(\theta - \theta')$. So for a random-walk candidate density, the acceptance probability simplifies to be

$$\begin{aligned}\alpha(\theta, \theta') &= min\left[1, \frac{g(\theta'|x)q(\theta', \theta)}{g(\theta|x)q(\theta, \theta')}\right] \\ &= min\left[1, \frac{g(\theta'|x)}{g(\theta|x)}\right]. \quad \text{See[2].}\end{aligned}$$

Single Parameter with an Independent Candidate Density

Hastings in (1970) introduced Markov chains with candidate generating density that did not depend on the current value of the chain. These are called independent candidate distribution

$$q(\theta, \theta') = q_2(\theta')$$

for some function $q_2(\theta)$.

For an independent candidate density, the acceptance probability simplifies to be

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

$$= min\left[1, \frac{g(\theta'|x)}{g(\theta|x)} \times \frac{q_2(\theta)}{q_2(\theta')}\right].$$

Example 4.3.1. [2] Suppose we have a posterior density given by

$$g(\theta|x) = .8 \times e^{-\frac{1}{2}\theta^2} + .2 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(\theta-3)^2}.$$

This is a mixture of a normal$(0, 1^2)$ and a normal$(3, 2^2)$.

Let us use the normal candidate density with variance $\sigma^2 = 1$ centered around the current value as our random-walk candidate density distribution. Its shape is given by

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

Let the starting value be $\theta = 2$. Since the random-walk candidate density is symmetric about the current value, the acceptance probability

$$\alpha = min\left[1, \frac{g(\theta'|x)}{g(\theta|x)}\right].$$

The Metropolis-Hastings algorithm proceeds as follows:

Let the starting value be $\theta^{(0)} = 2$.

Draw $\theta' = 1.767$ from $q(\theta^{(n-1)}, \theta') = q(\theta^{(0)}, \theta') = q(2, \theta') = e^{-\frac{1}{2}(\theta'-2)^2}$.

Calculate the probability $\alpha(2, \theta')$.

$$\begin{aligned}
\alpha(2, 1.767) &= min\left[1, \frac{g(\theta'|x)}{g(\theta|x)}\right] \\
&= min\left[1, \frac{g(1.767|x)}{g(2|x)}\right] \\
&= min\left[1, \frac{.8 \times e^{-\frac{1}{2}\times 1.767^2} + .2 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(1.767-3)^2}}{.8 \times e^{-\frac{1}{2}\times 2^2} + .2 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(2-3)^2}}\right] \\
&= min\left[1, 1.5470\right] \\
&= 1
\end{aligned}$$

Draw $u = .773$ from $U(0, 1)$.

Now $u < \alpha(2, 1.767)$ so let $\theta^{(1)} = 1.767$.

Now we start with $\theta^{(1)} = 1.767$.

Draw $\theta' = 1.975$ from $q(\theta^{(n-1)}, \theta') = q(\theta^{(1)}, \theta') = q(1.767, \theta') = e^{-\frac{1}{2}(\theta'-1.767)^2}$.

Calculate the probability $\alpha(1.767, \theta')$.

$$\alpha(1.767, 1.975) = min\left[1, \frac{g(1.975|x)}{g(1.767|x)}\right]$$

$$= min\left[1, \frac{.8 \times e^{-\frac{1}{2}\times 1.975^2} + .2 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(1.975-3)^2}}{.8 \times e^{-\frac{1}{2}\times 1.767^2} + .2 \times \frac{1}{2} e^{-\frac{1}{2\times 2^2}(1.767-3)^2}}\right]$$

$$= .804$$

Draw $u = .933$ from $U(0, 1)$.

Now $u > \alpha(1.767, 1.975)$ so let $\theta^{(2)} = \theta^{(1)} = 1.767$.

You can continue this process, the next table give a summary of first six draws of the chain using the random-walk candidate density.

Table 4.1: Summary of first six draws of the chain using the random-walk candidate density.

Draw	Current value	Candidate	α	u	Accept
1	2.000	1.767	1.000	.773	yes
2	1.767	1.975	.804	.933	no
3	1.767	.547	1.000	.720	yes
4	.547	1.134	.659	.240	yes
5	1.134	1.704	.553	.633	no
6	1.134	-.836	1.000	.748	yes

Therefore $\theta^{(n)} = 2.000, 1.767, 1.767, .547, 1.134, 1.134, \ldots$.

4.3.2 Metropolis-Hastings Algorithm for Multiple Parameters

Suppose we have p parameters $\theta_1, \ldots, \theta_p$. Let the parameter vector be

$$\Theta = (\theta_1, \ldots, \theta_p)$$

Let $q(\Theta', \Theta)$ be the candidate density when the chain is at Θ and let $g(\Theta|x)$ be the posterior density. The probability of moving in this case is given by

$$\alpha(\Theta, \Theta') = min\left[1, \frac{g(\Theta'|x)q(\Theta', \Theta)}{g(\Theta|x)q(\Theta, \Theta')}\right]. \tag{4.4}$$

You can use the same steps of Metropolis-Hastings algorithm which used in the case of single parameter.

Multiple Parameters with a Random-Walk Candidate Density

Since we are using a random-walk candidate density it given by

$$q(\Theta, \Theta') = q_1(\theta'_1 - \theta_1, \ldots, \theta'_p - \theta_p)$$

where the function $q_1(,\ldots,)$ is symmetric about 0. Thus we can write the candidate density as

$$q(\Theta, \Theta') = q_1(\Theta' - \Theta).$$

Because of the symmetry $q_1(\Theta' - \Theta) = q_1(\Theta - \Theta')$, so for a random-walk candidate density, the acceptance probability simplifies to be

$$\begin{aligned}\alpha(\Theta, \Theta') &= min\left[1, \frac{g(\Theta'|x)q(\Theta', \Theta)}{g(\Theta|x)q(\Theta, \Theta')}\right] \\ &= min\left[1, \frac{g(\Theta'|x)}{g(\Theta|x)}\right].\end{aligned}$$

Multiple Parameters with an Independent Candidate Density

When an independent candidate density is used for multiple parameters

$$q(\Theta, \Theta') = q_2(\Theta').$$

The acceptance probability for a chain that uses an independent candidate density simplifies to

$$\begin{aligned}\alpha(\Theta, \Theta') &= min\left[1, \frac{g(\Theta'|x)q(\Theta', \Theta)}{g(\Theta|x)q(\Theta, \Theta')}\right] \\ &= min\left[1, \frac{g(\Theta'|x)}{g(\Theta|x)} \times \frac{q_2(\Theta)}{q_2(\Theta')}\right].\end{aligned}$$

See [2].

4.3.3 Blockwise Metropolis-Hastings Algorithm

Let the parameter vector be partitioned into blocks

$$\Theta = \theta_1, \theta_2, \ldots, \theta_J$$

where θ_j is a block of parameters. Let θ_{-j} be all the other parameters not in block j.

Hastings in (1970) suggested that, instead of applying the **Metropolis-Hastings algorithm** to the whole parameter vector Θ all at once, that the algorithm be applied sequentially to each block of

parameters θ_j in turn, conditional on knowing the values of all other parameters not in that block.

Steps of Blockwise Metropolis-Hastings

1. Start at point in parameter space $\theta_1^{(0)}, \ldots, \theta_J^{(0)}$.
2. For $n = 1, \ldots, N$
    - For $j = 1, \ldots, J$
    - draw candidate from $q(\theta_j^{(n-1)}, \theta_j' | \theta_1^{(n)}, \ldots, \theta_{j-1}^{(n)}, \theta_{j+1}^{(n-1)}, \ldots, \theta_J^{(n-1)})$.
    - Calculate the acceptance probability $\alpha(\theta_j^{(n-1)}, \theta_j' | \theta_1^{(n)}, \ldots, \theta_{j-1}^{(n)}, \theta_{j+1}^{(n-1)}, \ldots, \theta_J^{(n-1)})$.
    - Draw u from $U(0, 1)$.
    - if $u < \alpha(\theta_j^{(n-1)}, \theta_j')$ then let $\theta_j^{(n)} = \theta_j'$, else let $\theta_j^{(n)} = \theta_j^{(n-1)}$.

At each step for each block in turn, we draw the candidate θ_j from the candidate density, calculate the acceptance probability $\alpha(\theta_j, \theta_j')$, and either move that block of parameters to that candidate θ_j', or keep that block at the current value θ_j, depending on whether or not a random draw from a *uniform*(0,1) random variable is less than the acceptance probability. [2].

4.4 Gibbs Sampling

Gibbs sampling algorithm was developed by Geman (1984) as a method for recreating images from a noisy signal. They named it after Josiah Willard Gibbs who had determined a similar algorithm could be used to determine the energy states of gasses at equilibrium. He would cycle through the particles, drawing each one conditional on the energy levels of all other particles. His algorithm became the basis for the field of statistical mechanics, and made This sparked a big increase in the use of Bayesian methods in applied statistics.

Gibbs sampling algorithm is just a special case of the blockwise Metropolis-Hastings algorithm, the case where we draw each candidate block from its true conditional density given all the other blocks. See [26].

4.4.1 Gibbs Sampling Procedure

First, let the parameter vector be partitioned into blocks

$$\Theta = \theta_1, \theta_2, \ldots, \theta_J$$

where θ_j is the j^{th} block of parameters. Each block contains one or more parameters.

Let θ_{-j} be the set of all the other parameters not in block j.

The proportional form of Bayes theorem,

$$g(\theta_1, \ldots, \theta_J | x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n | \theta_1, \ldots, \theta_J) \times g(\theta_1, \ldots, \theta_J)$$

gives the shape of the joint posterior density of all the parameters, where

$$f(x_1, \ldots, x_n | \theta_1, \ldots, \theta_J) \quad \text{and} \quad g(\theta_1, \ldots, \theta_J)$$

are the joint likelihood and the joint prior density for all the parameters. This gives us the shape of the joint posterior, not its scale.

Gibbs sampling requires that we know the full conditional distribution of each block of parameters θ_j, given all the other parameters θ_{-j} and the data $X = (x_1, \ldots, x_n)$.

Let the full conditional distribution of block θ_j be denoted

$$g(\theta_j | \theta_{-j}, X) = g(\theta_j | \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_J, X).$$

In Gibbs sampling, we will cycle through the parameter blocks in turn, drawing each one from its full conditional distribution given the most recent values of the other parameter blocks, and all the observed data.

Gibbs sampling is a special case of the blockwise Metropolis-Hastings algorithm, where the conditional candidate density for each block of parameters is the conditional density of that block, given all the parameters in the other blocks and the data. **Since the candidates are being drawn from the correct full conditional distribution, every draw will be accepted.** [2].

Steps of the Gibbs Sampler

1. At time $n = 0$ start from an arbitrary point in the parameter space $\theta^0 = (\theta_1^{(0)}, \ldots, \theta_J^{(0)})$. Note: usually the starting point is chosen by taking a random draw from the joint prior distribution of the parameters.

2. For $n = 1, \ldots, N$.

   - For $j = 1, \ldots, J$, draw $\theta_j^{(n)}$ from $g(\theta_j | \theta_1^{(n)}, \ldots, \theta_{j-1}^{(n)}, \theta_{j+1}^{(n-1)}, \ldots, \theta_J^{(n-1)}, X)$.

3. The stationary distribution of $\theta^{(N)} = (\theta_1^{(N)}, \ldots, \theta_J^{(N)})$ is the true posterior $g(\theta_1, \ldots, \theta_J | X)$. This means that for a large N the value $\theta^{(N)} = (\theta_1^{(N)}, \ldots, \theta_J^{(N)})$ will be approximately a random draw from the true posterior. See [2].

In the next chapter we interest at using the Gibbs sampler to do Bayesian inference on finite mixture of Poisson distributions. Our main task in this thesis depends on using the Gibbs sampler to simulate a Markov chain which has the posterior density of our mixture model as its stationary distribution. Then we use the resulting sample to make the suitable Bayesian computations and draw conclusion about the unknown parameters of the Poisson mixture model.

Chapter 5

Bayesian Analysis of Finite Poisson Mixtures

Poisson is a useful and widely used distribution, it plays an important role in modeling discrete count 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. Fortunately, when the population is heterogeneous, Poisson mixture instead of homogeneous Poisson can be built for the data.

Poisson mixture model is an important and flexible model. It plays a crucial role in many areas such as finance, biology, physics, sociology, agriculture and zoology because it can model discrete count data with heterogeneity and has the advantages that the homogeneous Poisson doesn't have, such as making more accurate estimates and hypothesis testing.

For example, if the China Mobile Communication Corporation wants to know how many signal towers should be built in Shanghai, then the number of calls income and outgo Shanghai should be estimated first. As the number of calls may have different means at different periods, for example there are usually more calls during 7 and 9 p.m., so the number of calls should be built with Poisson mixture instead of homogeneous Poisson. See [37].

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

sampler as one of Markov chain Monte Carlo (MCMC) methods to draw samples from the posterior of the Poisson mixtures in order to use them in the Bayesian analysis.

5.1 Finite Poisson Mixture Model

The probability function of the $k-$finite Poisson mixture is given by

$$f(x|\lambda, p) = \sum_{j=1}^{k} p_j \frac{e^{-\lambda_j}\lambda_j^x}{x!} \tag{5.1}$$

where $p = (p_1, p_2, ..., p_k)$, for some probabilities $p_j > 0$, $j = 1, ..., k$, $k > 1$, with $\sum_{j=1}^{k} p_j = 1$. $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$, and we assume that $\lambda_1 < \lambda_2 < \ldots < \lambda_k$ to ensure the identifiability of the above finite mixture. See [5].

Remember that identifiability means that if we have

$$f(x) = \sum_{j=1}^{k} p_j f(x|\lambda_j) \quad \text{and} \quad f'(x) = \sum_{j=1}^{k'} p'_j f(x|\lambda'_j),$$

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 $\theta_j = \theta'_{\pi_j}$. See [32].

5.1.1 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.

Firstly we introduce the missing data indicators z_i, $i = 1, 2, \ldots, n$.

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

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

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, $\sum_{j=1}^{k} z_{ij} = 1$. [8].

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 $j = 1, ..., k$ and $\sum_{j=1}^{k} p_j = 1$.
Thus the density $f(x_i|z_{ij} = 1)$ is $Poisson(\lambda_j)$, 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 $z_{ij} \sim Bernolli(p_j)$, and for each $z_i = (z_{ij})_{j=1}^{k}$, we have

$$z_i|p \sim multinomial(1, p_1, \ldots, p_k). \quad \text{See[3].}$$

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

$$f(z_{i1}, \ldots, z_{ik}|p_1, \ldots, p_k) = \frac{1!}{z_{i1}! \cdots z_{ik}!} \prod_{j=1}^{k} p_j^{z_{ij}} = \prod_{j=1}^{k} p_j^{z_{ij}}.$$

Since z_i's are independent of each others then, 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 $X = X_1, X_2, \ldots, X_n$ be an iid random sample drawn from a Poisson mixture density.
The likelihood density of the mixture is:

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

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

$$\begin{aligned} f(\mathbf{x}, z|\lambda, p) &= f(\mathbf{x}|z, \lambda, p) f(z|p, \lambda, \mathbf{x}) \\ &= f(\mathbf{x}|z, \lambda) f(z|p) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} \left(\frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \right)^{z_{ij}} \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j)^{z_{ij}} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} \left(p_j \frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \right)^{z_{ij}}. \end{aligned} \tag{5.2}$$

See[7].

5.1.2 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. See [34]).

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

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

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

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

  $$\mathrm{B}(\boldsymbol{\delta}) = \frac{\prod_{i=1}^{K} \Gamma(\delta_i)}{\Gamma\left(\sum_{i=1}^{K} \delta_i\right)}, \qquad \boldsymbol{\delta} = (\delta_1, \cdots, \delta_k). \quad \text{See[38].}$$

  Now if $\pi(p)$ denote the prior density of proportions p of our poisson mixture then, we will assume that, $p \sim Dir(p_1, \ldots, p_k, \delta_1, \cdots, \delta_k)$, with $\delta_j = 1, \forall j = 1, ..., k$.

  So the density of p is

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

  Note that we get the forth 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 multinomial parameter $p = (p_1, ..., p_k)$ is $Dirichlet(\delta)$.

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

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

where $x_{ij} \in \{0, 1\}$, and $\sum_{j=1}^k x_{ij} = 1$. [39].

Note that, we have a single trial results in exactly one of some fixed finite number k possible outcomes, with probabilities $p_1, ..., p_k$ (so that $p_j \in (0, 1)$ for $j = 1, ..., k$, and $\sum_{j=1}^k p_j = 1$). And suppose the prior distributed as $Dirichlet(\delta)$, that is, the prior density is given by

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

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:

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

The posterior density is:

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

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 $f(p|\mathbf{x})$ is in the same family as the prior density $f(p)$ with different parameters.

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

- For parameters $\lambda_j, j = 1, \ldots, k$ a gamma density is often chosen as a prior (as chosen by Viallefont, V. and others. See [34]). That is if $f(\lambda_j)$ denote the prior density of the j^{th} parameter of Poisson mixture then

$$\lambda_j \sim gamma(\alpha, \beta).$$

It is important to point out that, all $\lambda_j s$ have gammas densities but differ in the variables α, β. Next we will prove this for Poisson(θ) with parameter θ.

Let $X_1, X_2, ..., X_n$ be *i.i.d.* $Poisson(\theta)$, and suppose the prior density as $gamma(\alpha, \beta)$, that is, the prior density is given by

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

The likelihood density is:

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

The posterior density is:

$$\begin{aligned}
f(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta) f(\theta) \\
&= \frac{1}{x_1! x_2! ... x_n!} e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x_1! x_2! ... x_n!} \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-n\theta - \beta\theta} \\
&= \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x_1! x_2! ... x_n!}}_{\text{does not involve } \theta} \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-(n+\beta)\theta}.
\end{aligned}$$

We do not write the term which does not involve θ.

The posterior density becomes:

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

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

So,

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

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

5.1.3 The posterior density

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

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

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

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

$$\begin{aligned} f(\lambda, p, z|\mathbf{x}) &= \frac{f(\lambda, p, z, \mathbf{x})}{f(\mathbf{x})} \\ &\propto f(\lambda, p, z, \mathbf{x}) \\ &= f(\mathbf{x}, z|\lambda, p) g(\lambda)\pi(p) \\ &= f(\mathbf{x}|\lambda, z) f(z|p) g(\lambda)\pi(p) \qquad \text{(refer to Equations 5.2)} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}} \prod_{i=1}^{n}\prod_{j=1}^{k}(p_j)^{z_{ij}} \times g(\lambda_1) \times ... \times g(\lambda_k) \times (k-1)! \end{aligned}$$

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

5.2 Full Conditional Posterior Distributions

The Gibbs sampler is one of a set of Markov chain Monte Carlo (MCMC) methods, in which the full conditional posterior distributions of all parameters are required. Using our likelihood, priors, and the posterior joint density we obtain all full conditional posterior densities by ignoring all terms that are constant with respect to the parameter. Note that for our finite poisson mixture,
the likelihood distribution is

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

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

$$\lambda_j \sim gamma(\alpha, \beta)$$

for propotions p,

$$p \sim Dirichlet(\delta, \delta, ..., \delta), \delta = 1.$$

5.2.1 λ_j Posterior

The full conditional posterior density for λ_j is

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

Therefore

$$\lambda_j \sim gamma\left(\alpha + \sum_{i=1}^{n} x_i z_{ij}, \sum_{i=1}^{n} z_{ij} + \beta\right). \tag{5.3}$$

Note that the second proportionality because of that

$$\prod_{j=1}^{k}\left(\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}} = \underbrace{\left(\frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}\right)^{z_{i1}} \times ... \times \left(\frac{e^{-\lambda_{j-1}}\lambda_{j-1}^{x_i}}{x_i!}\right)^{z_{i(j-1)}}}_{\text{does not involve } \lambda_j} \times \left(\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}}$$
$$\times \underbrace{\left(\frac{e^{-\lambda_{j+1}}\lambda_{j+1}^{x_i}}{x_i!}\right)^{z_{i(j+1)}} \times ... \times \left(\frac{e^{-\lambda_k}\lambda_k^{x_i}}{x_i!}\right)^{z_{ik}}}_{\text{does not involve } \lambda_j}.$$

5.2.2 p Posterior

The full conditional posterior density for p is

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

It is clear that

$$p \sim Dirichlet\left(1 + \sum_{i=1}^{n} z_{i1}, \ldots, 1 + \sum_{i=1}^{n} z_{ik}\right). \tag{5.4}$$

5.2.3 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

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

Thus

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

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

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

5.2.4 Gibbs Updates for Fixed k

We consider a mixture of Poissons where, conditional on there being k components in the mixture. All the parameters of our Poisspn 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, $\ell = 1, \ldots, N$:

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

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

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

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

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

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

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

5.3 Bayesian Analysis on Poisson Mixture of Two Components

Consider a Poisson mixture of two components ($k = 2$) with the density function

$$f(x|\lambda_1, \lambda_2, p) = \sum_{j=1}^{2} p_j \frac{e^{-\lambda_j} \lambda_j^x}{x!} \tag{5.5}$$

where λ_1 the paramter of the first component, and λ_2 the parameter of the second component, and we assume that $\lambda_1 < \lambda_2$ to ensure the identifiability of the mixture.
$p = (p_1, p_2)$, for some probabilities $p_j > 0$, $j = \{1, 2\}$, with $\sum_{j=1}^{2} p_j = 1$. For each observation x_i, $i = 1, ..., n$ we have an indicator z_i such that

$$z_i = (z_{ij})_{j=1}^{2} = (z_{i1}, z_{i2})$$

where

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

Now $z_{ij} \sim Bernolli(p_j)$, and for each $z_i = (z_{ij})_{j=1}^{2}$, we have

$$z_i|p \sim multinomial(1, p_1, p_2).$$

So the density of the indicator $z_i = (z_{ij})_{j=1}^{2}$ is

$$f(z_{i1}, z_{i2}|p_1, p_2) = \prod_{j=1}^{2} p_j^{z_{ij}}.$$

Since z_i's are independent of each others then, the joint indicator density is

$$f(z|p) = \prod_{i=1}^{n} f(z_i|p) = \prod_{i=1}^{n} \prod_{j=1}^{2} p_j^{z_{ij}}, \quad \text{where } z = (z_1, ..., z_n).$$

Let $X_1, X_2, \ldots, X_n$ be iid random sample drawn from a Poisson mixture density of two components. By using the indicators z_i the likelihood density of the mixture is:

$$\begin{aligned} f(\mathbf{x}, z|\lambda_1, \lambda_2, p) &= f(\mathbf{x}|z, \lambda_1, \lambda_2, p) f(z|p, \lambda_1, \lambda_2, \mathbf{x}) \\ &= f(\mathbf{x}|z, \lambda_1, \lambda_2) f(z|p) \end{aligned}$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{2}\left(\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}} \prod_{i=1}^{n}\prod_{j=1}^{2}(p_j)^{z_{ij}} \tag{5.6}$$
$$= \prod_{i=1}^{n}\prod_{j=1}^{2}\left(p_j\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}}.$$

We assume that the parameters λ_1, λ_2 are independent and all priors are distributed independently of each other.

We will apply the general case of k components in the case of 2 components.

The prior densities on the parameters are:

$\lambda_1|\alpha_1, \beta_1 \sim gamma(\alpha_1, \beta_1)$, $\alpha_1, \beta_1 > 0$

$$g_1(\lambda_1) = \frac{{\beta_1}^{\alpha_1}}{\Gamma(\alpha_1)}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}$$

$\lambda_2|\alpha_2, \beta_2 \sim gamma(\alpha_2, \beta_2)$, $\alpha_2, \beta_2 > 0$

$$g_2(\lambda_2) = \frac{{\beta_2}^{\alpha_2}}{\Gamma(\alpha_2)}\lambda_2^{\alpha_2-1}e^{-\beta_2\lambda_2}$$

$p \sim Dirichlet(\delta, \delta), \delta = 1.$

$$\pi(p) = (2-1)! = 1$$

Inference for this model is based on the 4-dimensional posterior distribution $f(\lambda_1, \lambda_2, p, z|\mathbf{x})$ where $\mathbf{x} = (x_1, ..., x_n)$. The posterior density is obtained up to a multiplicative constant by multiplying the likelihood times the joint prior of the parameters. This gives

$$f(\lambda_1, \lambda_2, p, z|\mathbf{x}) \propto \prod_{i=1}^{n}\prod_{j=1}^{2}\left(\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}} \prod_{i=1}^{n}\prod_{j=1}^{2}(p_j)^{z_{ij}} \times g_1(\lambda_1)g_2(\lambda_2)\pi(p).$$

By using the above joint posterior density we can obtain full conditional distributions for each parameter by ignoring all terms that are constant with respect to the parameter.

- Full conditional for λ_1

$$f(\lambda_1|\lambda_2, p, z, \mathbf{x}) \propto \prod_{i=1}^{n}\prod_{j=1}^{2}\left(\frac{e^{-\lambda_j}\lambda_j^{x_i}}{x_i!}\right)^{z_{ij}} \times g_1(\lambda_1)$$

$$
\begin{aligned}
&\propto \prod_{i=1}^{n} \left(\frac{e^{-\lambda_1} \lambda_1^{x_i}}{x_i!} \right)^{z_{i1}} \frac{{\beta_1}^{\alpha_1}}{\Gamma(\alpha_1)} \lambda_1^{\alpha_1 - 1} e^{-\beta_1 \lambda_1} \\
&\propto \prod_{i=1}^{n} \left(e^{-\lambda_1} \lambda_1^{x_i} \right)^{z_{i1}} \times \lambda_1^{\alpha_1 - 1} e^{-\beta_1 \lambda_1} \\
&= e^{-\lambda_1 \sum_{i=1}^{n} z_{i1}} \times \lambda_1^{\sum_{i=1}^{n} x_i z_{i1}} \times \lambda_1^{\alpha_1 - 1} e^{-\beta_1 \lambda_1} \\
&= \left(e^{-\lambda_1 \sum_{i=1}^{n} z_{i1}} \times e^{-\beta_1 \lambda_1} \right) \left(\lambda_1^{\sum_{i=1}^{n} x_i z_{i1}} \times \lambda_1^{\alpha_1 - 1} \right) \\
&= \left(e^{-\lambda_1 (\sum_{i=1}^{n} z_{i1} + \beta_1)} \right) \left(\lambda_1^{\sum_{i=1}^{n} x_i z_{i1} + \alpha_1 - 1} \right) \\
&\propto gamma \left(\alpha_1 + \sum_{i=1}^{n} x_i z_{i1}, \sum_{i=1}^{n} z_{i1} + \beta_1 \right).
\end{aligned}
$$

- Full conditional for λ_2

$$
\begin{aligned}
f(\lambda_2 | \lambda_1, p, z, \mathbf{x}) &\propto \prod_{i=1}^{n} \prod_{j=1}^{2} \left(\frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \right)^{z_{ij}} \times g_2(\lambda_2) \\
&\propto \prod_{i=1}^{n} \left(\frac{e^{-\lambda_2} \lambda_2^{x_i}}{x_i!} \right)^{z_{i2}} \frac{{\beta_2}^{\alpha_2}}{\Gamma(\alpha_2)} \lambda_2^{\alpha_2 - 1} e^{-\beta_2 \lambda_2} \\
&\propto \prod_{i=1}^{n} \left(e^{-\lambda_2} \lambda_2^{x_i} \right)^{z_{i2}} \times \lambda_2^{\alpha_2 - 1} e^{-\beta_2 \lambda_2} \\
&= e^{-\lambda_2 \sum_{i=1}^{n} z_{i2}} \times \lambda_2^{\sum_{i=1}^{n} x_i z_{i2}} \times \lambda_2^{\alpha_2 - 1} e^{-\beta_2 \lambda_2} \\
&= \left(e^{-\lambda_2 \sum_{i=1}^{n} z_{i2}} \times e^{-\beta_2 \lambda_2} \right) \left(\lambda_2^{\sum_{i=1}^{n} x_i z_{i2}} \times \lambda_2^{\alpha_2 - 1} \right) \\
&= \left(e^{-\lambda_2 (\sum_{i=1}^{n} z_{i2} + \beta_2)} \right) \left(\lambda_2^{\sum_{i=1}^{n} x_i z_{i2} + \alpha_2 - 1} \right) \\
&\propto gamma \left(\alpha_2 + \sum_{i=1}^{n} x_i z_{i2}, \sum_{i=1}^{n} z_{i2} + \beta_2 \right).
\end{aligned}
$$

- Full conditional for p

$$
\begin{aligned}
f(p | z, \lambda_1, \lambda_2, x) &\propto \prod_{i=1}^{n} \prod_{j=1}^{2} (p_j)^{z_{ij}} \, \pi(p) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{2} (p_j)^{z_{ij}} \times 1 \\
&= \prod_{j=1}^{2} p_j^{\sum_{i=1}^{n} z_{ij}} \\
&= \prod_{j=1}^{2} p_j^{(\sum_{i=1}^{n} z_{ij} + 1) - 1}
\end{aligned}
$$

$$\propto Dirichlet\left(1+\sum_{i=1}^{n} z_{i1}, 1+\sum_{i=1}^{n} z_{i2}\right).$$

- Full conditional for z_i

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

  $$z_i = (z_{i1}, z_{i2}) = (z_{ij})_{j=1}^{2}$$

  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.

  Using Bayes' theorem we have,

  for fixed $i, i = 1, ..., n$, and for $j = \{1, 2\}$

  $$\begin{aligned} f(z_{ij} = 1|x_i, \lambda, p) &= \frac{f(x_i|\lambda, p, z_{ij} = 1) f(z_{ij} = 1|\lambda, p, x_i)}{\sum_{j=1}^{2} f(x_i|\lambda, p, z_{ij} = 1) f(z_{ij} = 1|\lambda, p, x_i)} \\ &= \frac{f(x_i|\lambda, z_{ij} = 1) f(z_{ij} = 1|p)}{\sum_{j=1}^{2} f(x_i|\lambda, z_{ij} = 1) f(z_{ij} = 1|p)} \\ &= \frac{f(x_i|\lambda_j) p_j}{\sum_{j=1}^{2} f(x_i|\lambda_j) p_j} \\ &= \frac{f(x_i|\lambda_j) p_j}{f(x_i)}. \end{aligned}$$

  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

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

  so

  $$z_i = (z_{i1}, z_{i2}) = (z_{ij})_{j=1}^{2} \sim multinomial(1, w_{i1}, w_{i2})$$

  where

  $$w_{ij} = \frac{f(x_i|\lambda_j) p_j}{f(x_i)}, \quad j = \{1, 2\}.$$

Remark 5.3.1. $\lambda_1, \lambda_2, z_i$, and p all 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.

5.3.1 Gibbs Updates

The Gibbs algorithm for the poisson mixture of two components is

Step 1: Pick a starting values of the parameters for the Markov chain, say $(\lambda_1^0, \lambda_2^0, p^0, z_1^0, ..., z_n^0)$

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

- **(a) Gibbs update of** λ_1: Sample λ_1^ℓ from $gamma\,(\alpha_1 + \sum_{i=1}^n x_i z_{i1}, \sum_{i=1}^n z_{i1} + \beta_1)$ using the most up-to-date values of z_{i1}.
- **(b) Gibbs update of** λ_2: Sample λ_2^ℓ from $gamma\,(\alpha_2 + \sum_{i=1}^n x_i z_{i2}, \sum_{i=1}^n z_{i2} + \beta_2)$ using the most up-to-date values of z_{i2}.
- **(c) Gibbs update of** p: Sample p^ℓ from $Dirichlet\,(1 + \sum_{i=1}^n z_{i1}, 1 + \sum_{i=1}^n z_{i2})$ using the most up-to-date values of z_{i1}, z_{i2}.
- **(d) Gibbs update of indicator** z_i: Sample z_i^ℓ from $multinomial(1, w_{i1}, w_{i2})$
  where $w_{ij} = \dfrac{f(x_i|\lambda_j)p_j}{f(x_i)}, \quad j = \{1, 2\}$
  using the most up-to-date values of λ_1, λ_2 and p.
- **(g)** We now have a new Markov chain state $({\lambda_1}^\ell, {\lambda_2}^\ell, p^\ell, {z_1}^\ell, ..., z_n^\ell)$.

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

5.4 Application

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

Our example uses a dataset from the Chandra Orion Ultradeep Project (COUP). This is a time series of X-ray emission from a flaring young star in the Orion Nebula Cluster. The raw data, which arrives approximately according to a Poisson process, gives the individual photon arrival times (in seconds) and their energies (in keV). The processed data we consider here is obtained by grouping the events into evenly-spaced time bins (10,000 seconds width). See [40].

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 Poisson 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.

5.4.1 Estimation results

- λ_1: mean = 4.77075, sd = 1.068456
- λ_2: mean = 10.61297, sd = 1.379152
- p_1: mean = 0.428015, sd = 0.1672902
- p_2: mean = 0.571985, sd = 0.1672902
- z_1: mean = 0.4253174, sd = 0.4943916
- z_2: mean = 0.5745826, sd = 0.4944067

5.4.2 Simulation Results

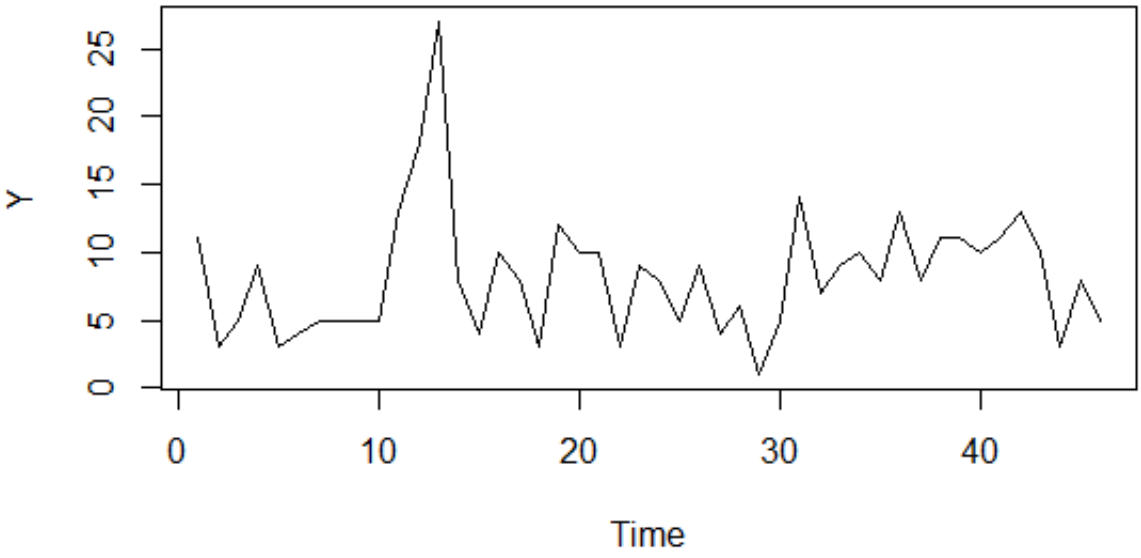

Figure 5.1: Time series plot for the data

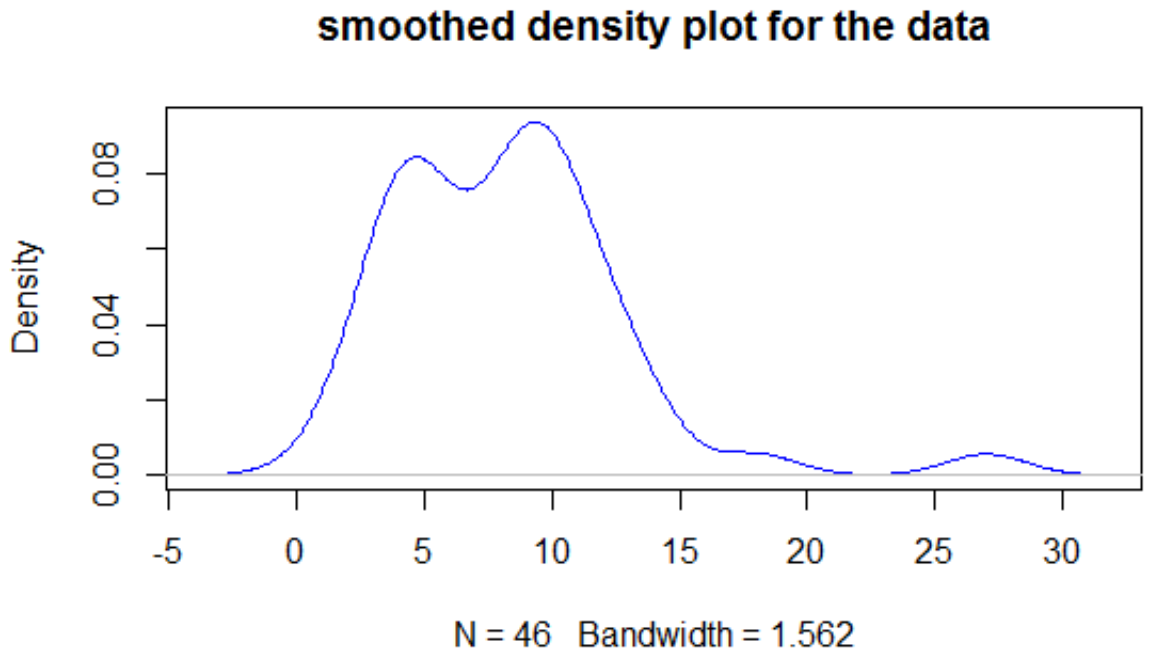

Figure 5.2: Smoothed density plot for the data

This figure show that the data is nonhomogeneous, so the best way to model this data is by use mixture model.

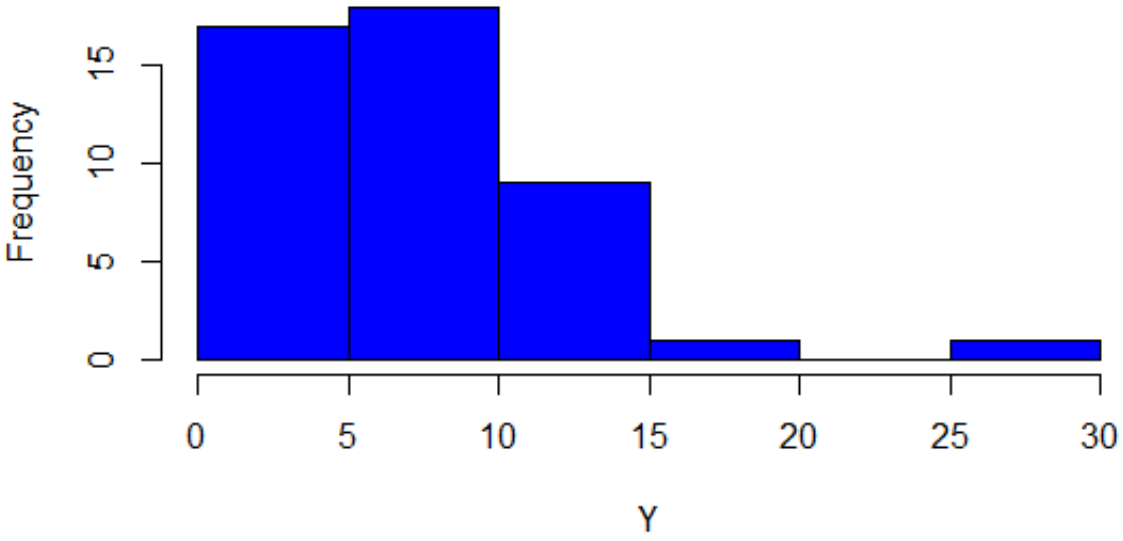

Figure 5.3: Histogram for the data

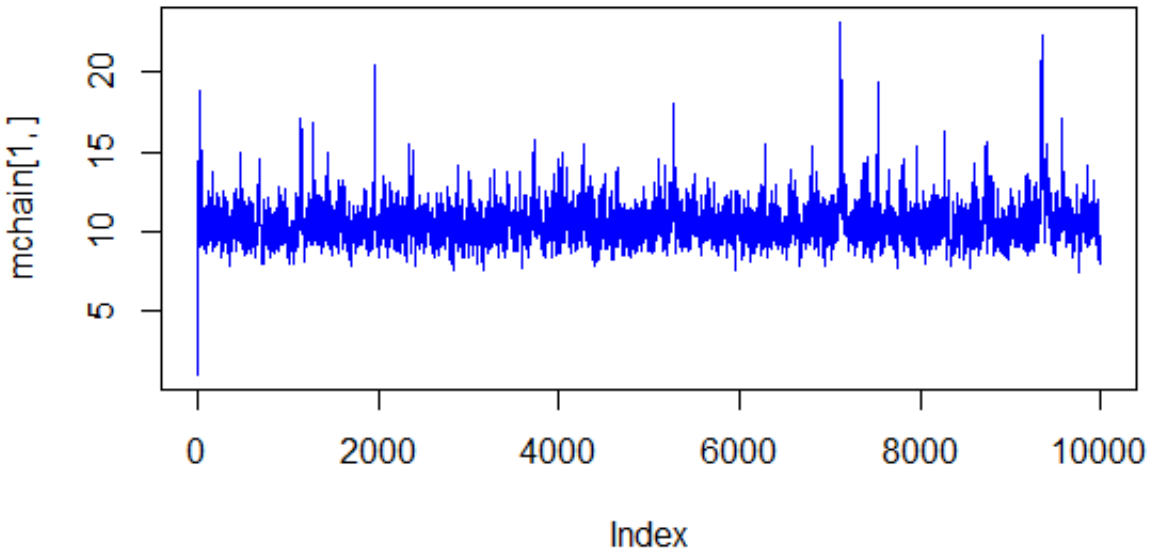

Figure 5.4: Markov Chain of λ_1

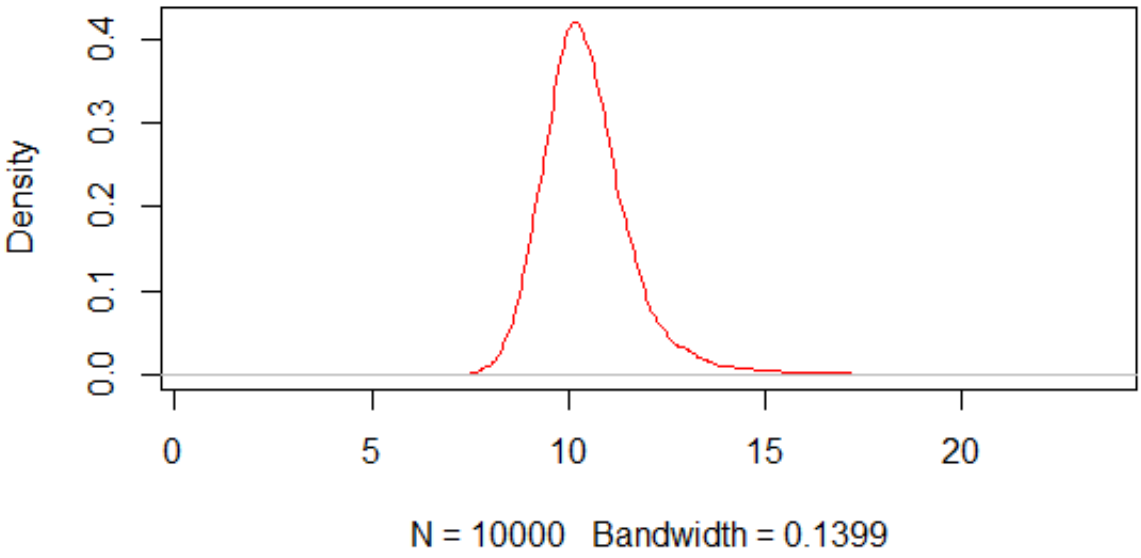

Figure 5.5: Density plot of λ_1

Clearly the density of λ_1 is gamma.

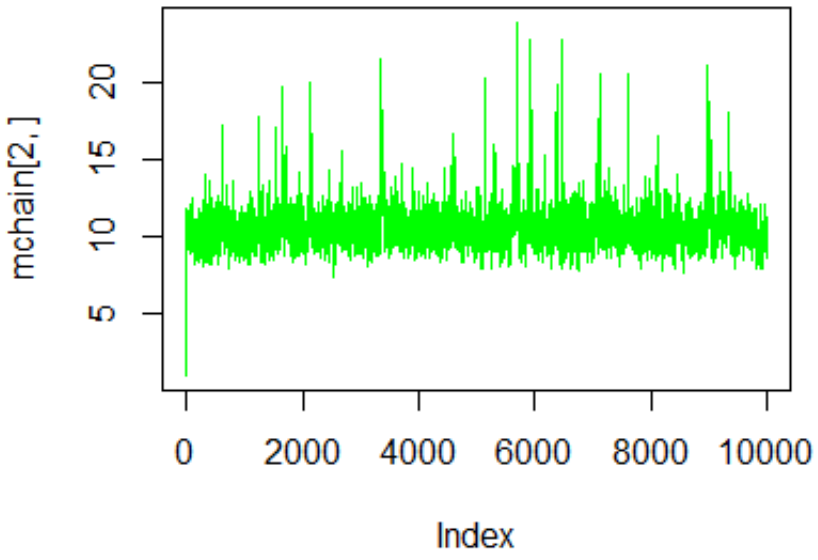

Figure 5.6: Markov Chain of λ_2

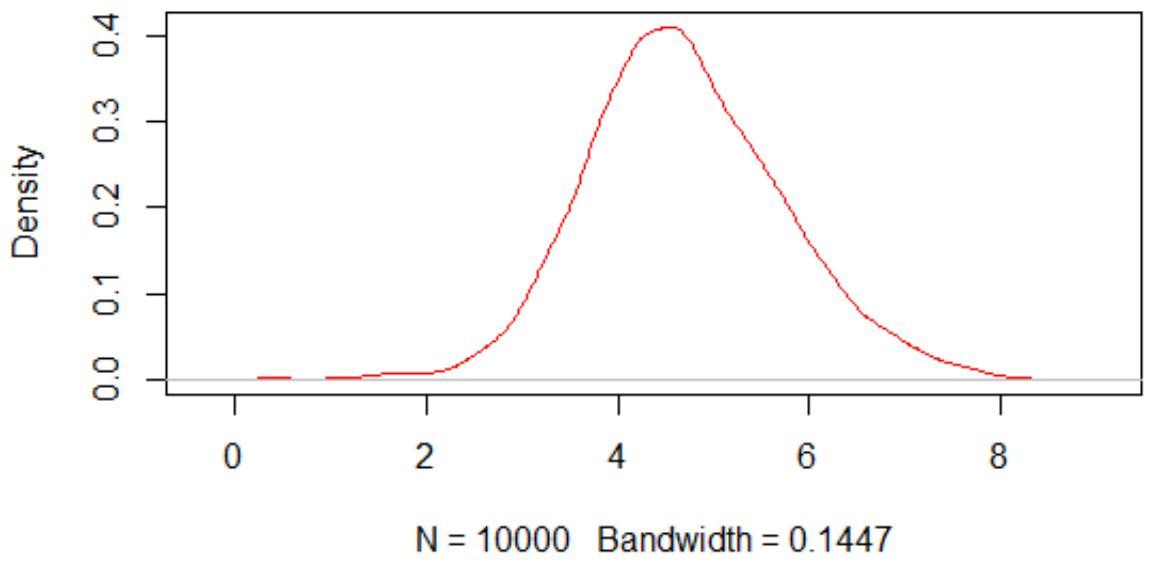

Figure 5.7: Density plot of λ_2

Its obvious that the density of λ_2 is gamma.

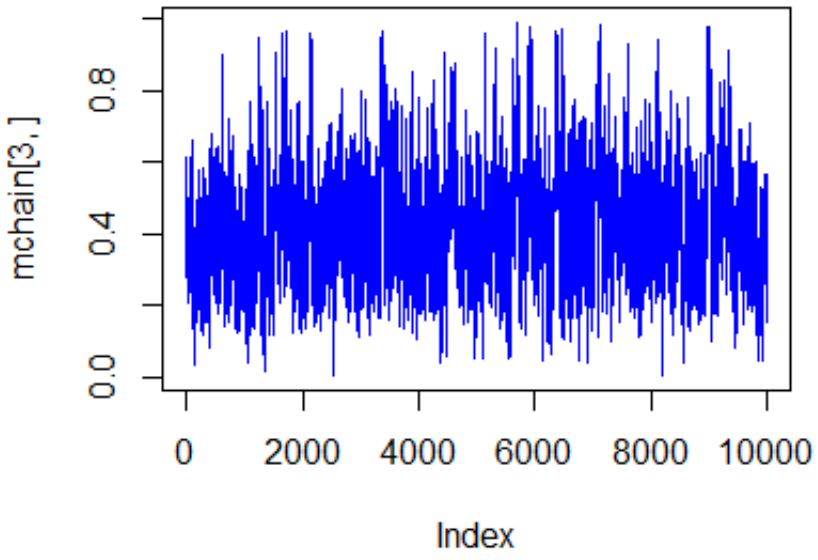

Figure 5.8: Markov Chain of p_1

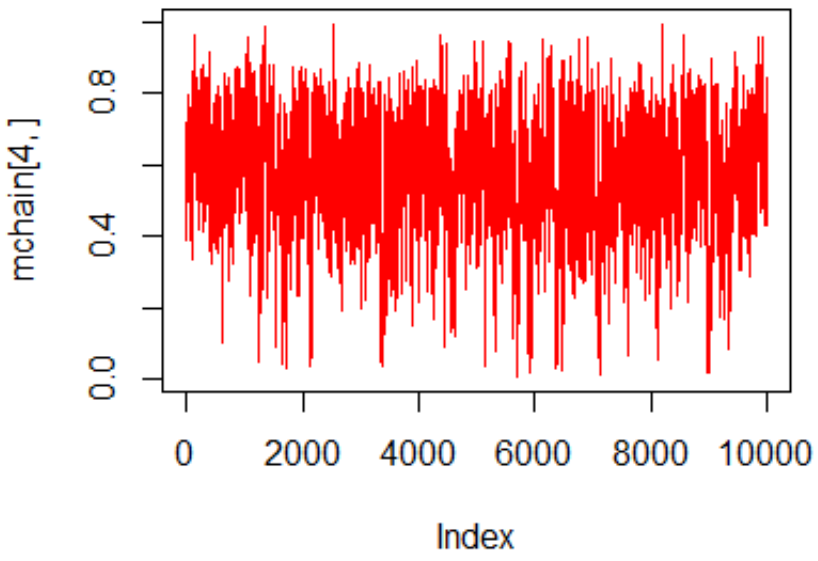

Figure 5.9: Markov Chain of p_2

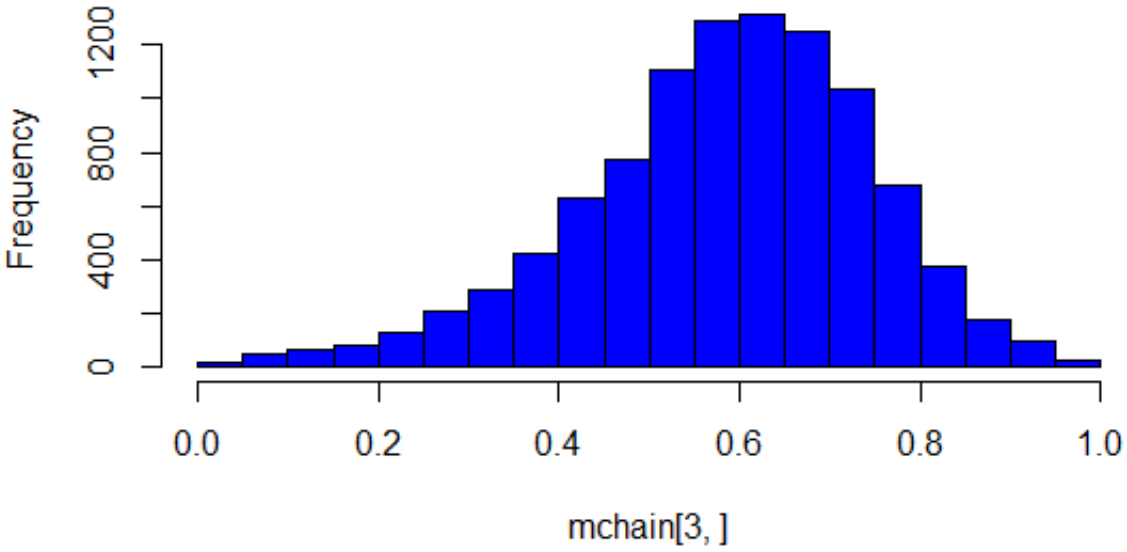

Figure 5.10: Histogram of p_1

Conclusion

- Bayesian statistics is based on a single tool, Bayes' theorem, which finds the posterior density of the parameters, given the data $f(\theta|\mathbf{x})$. It combines both the prior information we have given in the prior $f(\theta)$ and the information about the parameters contained in the observed data given in the likelihood $f(\mathbf{x}|\theta)$.

  We find the unsealed posterior by posterior proportional to prior times likelihood that is

  $$f(\theta|\mathbf{x}) \propto f(\theta) \times f(\mathbf{x}|\theta).$$

  The unsealed posterior has all the shape information. However, it 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.

- Computational Bayesian statistics is based on drawing a Monte Carlo random sample from the unsealed posterior. This replaces very difficult numerical calculations with the easier process of drawing random variables. Sometimes, particularly for high dimensional cases, this is the only feasible way to find the posterior.

- Mixed distributions are widely used to model data in which each observation is assumed to come from one of a number of distributions with different parameters.

- A Markov chain is a special type of stochastic process for which the future (the next step) depends only on the present state; it has no memory of how the present state was reached, this specific kind of "memorylessness" is called the Markov property.

- 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.
- Gibbs sampling algorithm is just a special case of the blockwise Metropolis-Hastings algorithm, the case where we draw each candidate block from its true conditional density given all the other blocks.
- 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.
- Poisson mixture model is an important and flexible model family. It plays a crucial role in many areas, because it can model discrete count data with heterogeneity and has the advantages that the homogeneous Poisson doesn't have, such as making more accurate estimates and hypothesis testing.

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] Marchini, J. (2008). *The Poisson Distribution*, University of Oxford, UK.

[21] 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.

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

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

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

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

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

[27] 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.

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

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

[30] 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.

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

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

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

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

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

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

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

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

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

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