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

The Gibbs Sampler in Bayesian Inference

Presented By
Mohammed S. Asfour

Supervisor
Professor Mohamed I. Riffi

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF MATHEMATICS
2014

Dedication

To...

my Parents;

my Wife;

my Brothers;

Zaher, Ashraf, Hazem, Omar and Anas.

my Sons;

Hamza and Hazem.

my Daughter;

Raghad.

Table of contents

List of Figures

List of Tables

‘

List of Symbols

$\mathbb{Z}$	*integers.*
$\mathbb{R}$	*real numbers.*
P	*probability measure.*
$\hat{\theta}$	*estimator of a parameter* θ.
$L(\theta, a)$	*loss function.*
$I(\theta)$	*Fisher information*
log	*logarithm.*
B_{01}	*Bayes factor.*
M_i	*model* i.
H_0	*null hypothesis.*
H_1	*alternative hypothesis.*
T	*transition matrix.*
$p(i, j)$	*probability of moving from the state* i *to the state* j *in one step.*
Ω	*state space.*
p_i	*probability mass function of* X_0.
$A(i)$	*set of all accessible states from* i.
T_i	*time for first visit to* i *given* $X_0 = i$.
f_{ii}^n	*probability of first recurrence to* i *at the* n^{th} *step.*
f_i	*probability of recurrence to* i.
π	*stationary distribution.*
x^*	*candidate point.*

List of Abbreviations

i.i.d.	*independent identically distributed.*
MC	*Markov Chain.*
MCMC	*Markov Chain Monte Carlo.*
gcd	*greatest common divisor.*

Acknowledgements

First of all, praise is to God who gives me the chance to complete my Master Degree.

I am grateful to my supervisor, Professor Mohamed I. Riffi, for a substantial amount of guidance, advice and support in producing this thesis.

I would like to thank Professor Abdallah Elhabeel and Dr Raed Salha for their enlightening discussions and all valuable suggestions, and all members of Mathematics Department in the Islamic University of Gaza for their help.

I am also thankful to all my brothers for their love and encouragement.

Finally, I would like to express my deep thanks to my family for their encouragement and support.

;

بسم الله الرحمن الرحيم

الملخص

في هذه الرسالة، سندرس طريقة معاين جيبس في الاستدلال البايزي لتوزيع له دالة كثافة مجهولة. طريقة معاين جيبس هي احدي طرق سلسلة ماركوف مونت كارلو.

طريقة معاين جيبس تستخدم بشكل واسع في الإحصاء و في مجالات أخري للتعرف علي التوزيعات الاحتمالية المعقدة ، وهذا يتم عن طريق استخدام اسلوب المحاكاة لبناء سلسلة ماركوف بحيث يكون توزيع الاتزان لسلسلة ماركوف يساوي التوزيع ألبعدي، بهذه الطريقة يمكن سحب عينة من التوزيع ألبعدي، واستخدام التوزيع ألبعدي في التعرف علي خصائص التوزيع المجهول.

بعد دراسة وافية لمعاين جيبس و عرض نظريات متعلقة بمعاين جيبس و بعض الأمثلة سوف نطبق طريقة معاين جيبس لدراسة حالة تتضمن عمل استدلال البايزي لتوزيع له دالة كثافة مجهولة.

Abstract

In this thesis, we will study the role of the Gibbs sampler in Bayesian inference about a complicated distribution. This complicated distribution has either an unknown probability density function (pdf) or a pdf that is only known up to a normalizing constant. The Gibbs sampler is one of the basic methods of the Markov Chain Monte Carlo (MCMC) algorithms.

It is widely used in statistics and other disciplines for exploring complicated probability distributions. This can be achieved by using simulation techniques in order to construct a Markov chain that has a stationary distribution equal to the pdf of the complicated distribution. In this way, we can draw a random sample from the posterior distribution and use it to draw conclusions about certain properties of the complicated distribution.

After thoroughly discussing the Gibbs sampler algorithm and present the relevant theorems and some examples, we will apply the Gibbs sampler technique to a case study that involves the making Bayesian inference about the underlying distribution.

Introduction

The main purpose of this thesis is to give a detailed introduction to the general idea of MCMC methods and apply the role of the Gibbs sampler in Bayesian inference about a complicated distribution. This thesis is organized in four chapters.
In Chapter 1, we give the basic definitions and some theorems from the Probability Theory which we need in this thesis. Then we introduce the definition of the posterior distribution and give some examples. This chapter contains the following sections:

- 1.1 Posterior Distributions.
- 1.2 Bayes Estimators.

In Chapter 2, we explore some types of priors, predictive distributions and model choice. This chapter contains the following sections:

- 2.1 Priors.
- 2.2 Predictive Distributions.
- 2.3 Model Choice.

In Chapter 3, we study some important defnitions and properties of Markov chains states, and we also study multistep transition probabilities, classification of Markov chain states and stationary distribution of Markov chains. This chapter contains the following sections:

- 3.1 Defnitions and basic properties.
- 3.2 Multistep transition probabilities.
- 3.3 Classifcation of Markov chain states.
- 3.4 Stationary distribution.
- 3.4 Detailed Balance condition.

In Chapter 4, first, we present some roles of Markov Chain Monte Carlo methods. Second, we introduce simulators which have used in simulate from posterior distributions by MCMC methods and give some examples. This chapter contains the following sections:

- 4.1 Markov Chain Monte Carlo Algorithms.
  - Metropolis-Hastings Algorithm.
  - Gibbs Sampler.
- 4.2 Simulation.

Chapter 1

Bayesian Statistics

1.1 Introduction

In this chapter, we introduce the basics of Bayesian statistics. Bayesian statistical analysis is concerned with calculating probability distributions of parameters in statistical models, where this statistical model describes the relationship between the parameters and the data in a mathematical model. Bayesian statistical analysis treats the parameters as random variables. A non-Bayesian statistical analysis treats parameters as fixed values without distribution. The first component of Bayesian statistical analysis is the prior distribution. The prior distribution is the distribution of the parameters befor any data is observed.

The second component of Bayesian statistical analysis is the sampling (likelihood) distribution. The sampling distribution is the distribution of the observed data conditioned on its parameters.

The last component of Bayesian statistical analysis is the posterior distribution. This distribution is calculated by combining the prior distribution and sampling distribution. We can use the posterior distribution to draw conclusions about certain properties of the population and compute point estimates of parameters.

1.2 Posterior Distributions

In this section, we introduce quick revision for the basic definitions and some theorems from the Probability Theory which we need in this thesis. Then, we introduce the definition of the posterior distribution and give some examples.

Axiom 1.2.1 (Probability Axioms)**.** [23] *Let S be the sample space of a random variable. Suppose that to each event A of S, a number denoted $P(A)$ is associated with A. If P satisfies the following axioms, then it is called a probability and the number $P(A)$ is said to be the probability of A.*

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

Axiom 2. $P(S) = 1$.

Axiom 3. *If $\{A_1, A_2, A_3, ...\}$ is a sequence of events, and $A_i \cap A_j = \phi$, $\forall i \neq j$, then*

$$P(\cup_{i=1}^{\infty} A_i) \quad = \quad \Sigma_{i=1}^{\infty} P(A_i).$$

Definition 1.2.1 (Conditional Probability)**.** [20] Let A and B be two events with $P(A) > 0$ and $P(B) > 0$. Then the conditional probability of A given B is:

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

Proposition 1.2.1 (Multiplication Rule)**.** [24] *Let A and B be two events with $P(A) > 0$ and $P(B) > 0$. Then*

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

Theorem 1.2.2 (Law of Total Probability)**.** [23] *If $\{B_1, B_2, ..., B_n\}$ is a sequence of mutually exclusive events of S such that $\bigcup_{i=1}^{n} B_i = S$ and $P(B_i) > 0$ for $i = 1, 2, ..., n$, then for any event A of S,*

$$P(A) \quad = \quad \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{1.3}$$

Bayes Theorem can now be stated, following immediately from the definition of conditional probability:

Theorem 1.2.3 (Bayes Theorem)**.** [24] *Let A and B be two events with $P(A) > 0$ and $P(B) > 0$. Then*

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

Theorem 1.2.4 (Bayes Theorem for Multiple Discrete Events)**.** [20] *Suppose that $E, F_1, ..., F_n$ are events from a sample space Ω, and that $P : \Omega \to [0, 1]$ is a probability distribution on Ω. Suppose that $\bigcup_{i=1}^{n} F_i = \Omega$, and that $F_i \cap F_j = \phi$, $\forall\ i \neq j$. Suppose $P(E) > 0$, $P(F_j) > 0$, $\forall j = 1, 2, ..., n$. Then*

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\Sigma_{i=1}^{n} P(E|F_i)P(F_i)},\ \forall j = 1, 2, ..., n. \tag{1.5}$$

Theorem 1.2.5 (Bayes Theorem for Continuous Parameters)**.** [24] *Suppose that two continuous random variables X and θ are given with pdf's $f(x|\theta)$ and $f(\theta)$. Then*

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

Table 1.1 contains the definitions of the distributions used in this thesis.

Table 1.1: Notation for common pdf's and pmf's

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

Notations. Suppose X_i is a random variable and x_i is its value, $i = 1, 2, ..., n$, and Θ is the space of θ.

- $\mathbf{x} = (x_1, x_2, ..., x_n)$.
- θ = unkown parameter.
- $f(x_1, x_2, ..., x_n|\theta)$ = likelihood distribution.

- $f(\theta)$ = prior distribution.

Applying Theorem 1.2.5, we have

$$f(\theta|x_1, x_2, ..., x_n) = \frac{f(x_1, x_2, ..., x_n|\theta)f(\theta)}{\int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n|\theta)f(\theta)d\theta}. \quad (1.7)$$

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

Let

$$z = \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n|\theta)f(\theta)d\theta.$$

Equation (1.7) becomes:

$$\begin{aligned} f(\theta|x_1, x_2, ..., x_n) &= \frac{f(x_1, x_2, ..., x_n|\theta)f(\theta)}{z} \\ &\propto f(x_1, x_2, ..., x_n|\theta)f(\theta). \end{aligned}$$

Therefore Bayes rule connects the posterior $f(\theta|x_1, x_2, ..., x_n)$ with the prior $f(\theta)$ via the formula:

$$f(\theta|x_1, x_2, ..., x_n) \propto f(x_1, x_2, ..., x_n|\theta)f(\theta), \quad (1.8)$$

in words:

$$Posterior \propto Likelihood \times Prior. \quad (1.9)$$

Note that we only write down the terms involving θ from (1.8). We do not care about the other terms which do not involve θ, because these are canceled in the ratio of equation (1.7).
Then we are able to identify the posterior distribution of θ just by looking at the numerator.

Example 1.2.1. *Let $X_1, X_2, ..., X_n$ be i.i.d. $Bi(n, \theta)$ with density*

$$f(x|\theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}, n \in \{1, 2, ...\}, x \in \{0, 1, ..., n\}, \theta \in (0, 1).$$

Also suppose the prior density is given by:

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

The likelihood density is:

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} \begin{pmatrix} n \\ x_i \end{pmatrix} \theta^{x_i}(1-\theta)^{n-x_i} \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} n \\ x_i \end{pmatrix} \right\} \theta^{x_1}\theta^{x_2}...\theta^{x_n}(1-\theta)^{n-x_1}(1-\theta)^{n-x_2}...(1-\theta)^{n-x_n} \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} n \\ x_i \end{pmatrix} \right\} \theta^{\sum_{i=1}^n x_i}(1-\theta)^{nn-\sum_{i=1}^n x_i} \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} n \\ x_i \end{pmatrix} \right\} \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n^2-\sum_{i=1}^n x_i}. \end{aligned}$$

The posterior density is:

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta) f(\theta) \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} n \\ x_i \end{pmatrix} \right\} \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n^2-\sum_{i=1}^n x_i} \frac{1}{beta(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \underbrace{\frac{1}{beta(\alpha, \beta)} \left\{ \prod_{i=1}^{n} \begin{pmatrix} n \\ x_i \end{pmatrix} \right\}}_{does\, not\, involve\, \theta} \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n^2-\sum_{i=1}^n x_i+\beta-1}. \end{aligned}$$

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

So, the last equation becomes:

$$f(\theta|\mathbf{x}) \quad \propto \quad \theta^{\sum_{i=1}^{n} x_i+\alpha-1}(1-\theta)^{n^2-\sum_{i=1}^{n} x_i+\beta-1}.$$

Clearly this is the density of a beta distribution with parameters: $\sum_{i=1}^{n} x_i + \alpha$ *and* $n^2 - \sum_{i=1}^{n} x_i + \beta$.

Therefore the posterior density becomes:

$$f(\theta|\mathbf{x}) = Be\left(\ \Sigma_{i=1}^{n} x_i + \alpha, n^2 + \beta - \Sigma_{i=1}^{n} x_i \ \right).$$

Example 1.2.2. [15] *Suppose* $X_1, X_2, ..., X_n$ *be an i.i.d. sample from the distribution with density*

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

Suppose the prior density for θ *is given by:*

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

The likelihood density is:

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

Therefore

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

The posterior density is:

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

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

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

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

$$f(\theta|\mathbf{x}) = Ga\left(n+1, \Sigma_{i=1}^n x_i + \mu \right).$$

1.3 Bayes Estimators

Suppose that $f(\theta|\mathbf{x})$ is the posterior distribution and $\mathbf{a}$ is our guess for θ. Note that $\mathbf{a}$ should be a function of $\mathbf{x} = (x_1, x_2, ..., x_n)$. We need to choose a good estimator $\hat{\theta}$ for the parameter θ, so, let us define $L(\theta, \mathbf{a})$ as the loss function.
The loss function has following properties:

1. $L(\theta, \mathbf{a}) \geq 0$.
2. $L(\theta, \mathbf{a})$ increases as the the distance between $\mathbf{a}$ and θ increases.
3. $L(\theta, \mathbf{a})$ is minimum if $\theta = \mathbf{a}$.

See [24].

Definition 1.3.1 (Posterior expected loss)**.** [24] Given a posterior density for θ, $f(\theta|\mathbf{x})$, the posterior expected loss of $\mathbf{a}$ is:

$$\varrho(f(\theta|\mathbf{x}), \mathbf{a}) \;=\; \int_{\Theta} L(\theta, \mathbf{a}) f(\theta|\mathbf{x}) d\theta.$$

We should choose that value of $\mathbf{a}$ which minimizes $\varrho(f(\theta|\mathbf{x}), \mathbf{a})$. The minimizer $\hat{\theta}$ is called Bayes estimator.

1.3.1 Squared Loss Function

In this subsection, we introduce the definition of the squared loss function and find the Bayes estimator of a parameter θ under squared loss function. Finally, we give an important example.

Definition 1.3.2 (Squared loss)**.** [24]
If $\theta \in \Theta$ is a parameter, and $\hat{\theta}$ is an estimate of θ, then $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ is called the squared loss function.

Theorem 1.3.1 (Posterior mean as a Bayes estimate under quadratic loss)**.** [24] *Under squared loss the Bayes estimator of θ is the mean of the posterior distribution, i.e.*

$$\hat{\theta} \;=\; E(\theta|\mathbf{x}) = \int_{\Theta} \theta f(\theta|\mathbf{x}) d\theta.$$

Proof. The posterior expected loss is:

$$\begin{aligned}
\varrho(f(\theta|\mathbf{x}), \hat{\theta}) &= \int_{\Theta} L(\theta, \hat{\theta}) f(\theta|\mathbf{x}) d\theta \\
&= \int_{\Theta} (\theta - \hat{\theta})^2 f(\theta|\mathbf{x}) d\theta \\
&= \int_{\Theta} (\theta^2 - 2\hat{\theta}\theta + \hat{\theta}^2) f(\theta|\mathbf{x}) d\theta \\
&= \int_{\Theta} \theta^2 f(\theta|\mathbf{x}) d\theta - 2\hat{\theta} \int_{\Theta} \theta f(\theta|\mathbf{x}) d\theta + \hat{\theta}^2 \int_{\Theta} f(\theta|\mathbf{x}) d\theta \\
&= \int_{\Theta} \theta^2 f(\theta|\mathbf{x}) d\theta - 2\hat{\theta} E(\theta|\mathbf{x}) + \hat{\theta}^2.
\end{aligned}$$

Differenate with respect to $\hat{\theta}$:

$$\varrho'(f(\theta|\mathbf{x}),\hat{\theta}) \;=\; -2E(\theta|\mathbf{x})+2\hat{\theta}.$$

Now set $\varrho'(f(\theta|\mathbf{x}),\hat{\theta})=0$, and solve it for $\hat{\theta}$.

$$-2E(\theta|\mathbf{x})+2\hat{\theta}=0,$$

$$\hat{\theta}=E(\theta|\mathbf{x}).$$

Then the Bayes estimator of θ is the mean of the posterior distribution. □

Example 1.3.1. [9] *Let $X_1, X_2, ..., X_n$ be i.i.d. $Pn(\theta)$.*

Also suppose the prior density is given by:

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

The likelihood density is:

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

The posterior density is:

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

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

The posterior density becomes:

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

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

$1+\Sigma_{i=1}^{n}x_i$ *and* $n+1$.

$$f(\theta|\mathbf{x}) \quad = \quad Ga(1+\Sigma_{i=1}^{n}x_i, n+1).$$

Hence the Bayes estimator of θ under squared loss is the mean of the posterior distribution

$$\hat{\theta} \quad = \quad \frac{1+\Sigma_{i=1}^{n}x_i}{n+1}.$$

1.3.2 Absolute Loss Function

In this subsection, we introduce the definition of the absolute loss function and find the Bayes estimator of a parameter θ under absolute loss function. Finally we give an important example.

Definition 1.3.3 (Absolute loss)**.** [24] If θ is a parameter, and $\hat{\theta}$ is an estimator of θ, then the Absolute loss function is:

$$L(\theta,\hat{\theta}) \quad = \quad |\theta-\hat{\theta}|.$$

Theorem 1.3.2 (Posterior median as a Bayes estimate under linear loss)**.** [24] *Under absolute loss, the Bayes estimator of θ is the median of the posterior density $f(\theta|\mathbf{x})$.*

Proof. The posterior expected loss is:

$$\begin{aligned}
\varrho(f(\theta|\mathbf{x}),\hat{\theta}) &= \int_{\Theta} L(\theta,\hat{\theta}) f(\theta|\mathbf{x}) d\theta \\
&= \int_{\Theta} |\theta-\hat{\theta}| f(\theta|\mathbf{x}) d\theta \\
&= \int_{\theta\geq\hat{\theta}} (\theta-\hat{\theta}) f(\theta|\mathbf{x}) d\theta - \int_{\theta<\hat{\theta}} (\theta-\hat{\theta}) f(\theta|\mathbf{x}) d\theta \\
&= \int_{\theta\geq\hat{\theta}} \theta f(\theta|\mathbf{x}) d\theta - \hat{\theta}\int_{\theta\geq\hat{\theta}} f(\theta|\mathbf{x}) d\theta + \hat{\theta}\int_{\theta<\hat{\theta}} f(\theta|\mathbf{x}) d\theta - \int_{\theta<\hat{\theta}} \theta f(\theta|\mathbf{x}) d\theta.
\end{aligned}$$

Differentiating this expression with respect to $\hat{\theta}$:

$$\varrho'(f(\theta|\mathbf{x}),\hat{\theta}) = -\int_{\theta\geq\hat{\theta}} f(\theta|\mathbf{x}) d\theta + \int_{\theta<\hat{\theta}} f(\theta|\mathbf{x}) d\theta.$$

Setting $\varrho'(f(\theta|\mathbf{x}),\hat{\theta}) = 0$ and solving it.

$$-\textstyle\int_{\theta\geq\hat{\theta}} f(\theta|\mathbf{x}) d\theta + \int_{\theta<\hat{\theta}} f(\theta|\mathbf{x}) d\theta = 0,$$

$$\textstyle\int_{\theta<\hat{\theta}} f(\theta|\mathbf{x}) d\theta = \int_{\theta\geq\hat{\theta}} f(\theta|\mathbf{x}) d\theta.$$

$\hat{\theta}$ = median.

Then the Bayes estimator of θ is the median of the posterior density $f(\theta|\mathbf{x})$. □

The Bayes estimator calculating process is difficult when the posterior distribution is not symmetric, in this case the Bayes estimator $\hat{\theta}$ of θ:

$$\hat{\theta} = median\, of\, the\, posterior.$$

Bayes estimator calculating process is easier than calculating it in previous case when the posterior distribution is symmetric, in this case the Bayes estimator $\hat{\theta}$ of θ:

$$\begin{aligned}
\hat{\theta} &= mean\, of\, the\, posterior \\
&= median\, of\, the\, posterior.
\end{aligned}$$

Example 1.3.2. [15] *Suppose $X_1, X_2, ..., X_n$ is i.i.d. $N(\theta, \sigma^2)$ where σ^2 is known. Let the prior density for θ be given by*

$$f(\theta) = \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}},$$

for known μ and τ^2.

The likelihood density is:

$$\begin{aligned}
f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \\
&= \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \frac{1}{\sigma\sqrt{2\pi}} ... \frac{1}{\sigma\sqrt{2\pi}}}_{n\text{-}copies} e^{-\frac{(x_1-\theta)^2}{2\sigma^2}} e^{-\frac{(x_2-\theta)^2}{2\sigma^2}} ... e^{-\frac{(x_n-\theta)^2}{2\sigma^2}} \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{(x_1-\theta)^2}{2\sigma^2} - \frac{(x_2-\theta)^2}{2\sigma^2} ... - \frac{(x_n-\theta)^2}{2\sigma^2}} \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}[(x_1-\theta)^2 + (x_2-\theta)^2 + ... + (x_n-\theta)^2]} \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2}.
\end{aligned}$$

The posterior density is:

$$\begin{aligned}
f(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta) f(\theta) \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{1}{2}((\Sigma_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}) + \frac{(\theta-\mu)^2}{\tau^2})} \\
&\propto e^{-\frac{1}{2}((\Sigma_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}) + \frac{(\theta-\mu)^2}{\tau^2})}.
\end{aligned}$$

Now let

$$M = \left(\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\sigma^2} \right) + \frac{(\theta-\mu)^2}{\tau^2},$$

then

$$f(\theta|\mathbf{x}) \propto e^{\frac{-M}{2}}.$$

We want to simplify M:

$$\begin{aligned}
M &= \left(\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\sigma^2} \right) + \frac{(\theta-\mu)^2}{\tau^2} \\
&= \sum_{i=1}^{n} \frac{{x_i}^2 - 2x_i\theta + \theta^2}{\sigma^2} + \frac{\theta^2 - 2\mu\theta + \mu^2}{\tau^2} \\
&= \sum_{i=1}^{n} \frac{{x_i}^2}{\sigma^2} - \sum_{i=1}^{n} \frac{2x_i\theta}{\sigma^2} + \sum_{i=1}^{n} \frac{\theta^2}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\mu\theta}{\tau^2} + \frac{\mu^2}{\tau^2} \\
&= \sum_{i=1}^{n} \frac{{x_i}^2}{\sigma^2} - \frac{2\theta \sum_{i=1}^{n} x_i}{\sigma^2} + \frac{n\theta^2}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\mu\theta}{\tau^2} + \frac{\mu^2}{\tau^2} \\
&= \frac{n\theta^2}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\theta \sum_{i=1}^{n} x_i}{\sigma^2} - \frac{2\mu\theta}{\tau^2} + \sum_{i=1}^{n} \frac{{x_i}^2}{\sigma^2} + \frac{\mu^2}{\tau^2} \\
&= \theta^2 \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) - 2\theta \left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu}{\tau^2} \right) + \sum_{i=1}^{n} \frac{{x_i}^2}{\sigma^2} + \frac{\mu^2}{\tau^2}.
\end{aligned}$$

Let

$$\begin{aligned}
a &= \frac{n}{\sigma^2} + \frac{1}{\tau^2}, \\
b &= \frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu}{\tau^2}, \\
c &= \sum_{i=1}^{n} \frac{{x_i}^2}{\sigma^2} + \frac{\mu^2}{\tau^2}.
\end{aligned}$$

Then

$$M = a\theta^2 - 2b\theta + c.$$

Notice that: a, b and c do not involve θ.

$$\begin{aligned}
M &= a\theta^2 - 2b\theta + c \\
&= a \left(\theta^2 - \frac{2\theta b}{a} \right) + c \\
&= a \left(\theta^2 - \frac{2\theta b}{a} + \frac{b^2}{a^2} - \frac{b^2}{a^2} \right) + c \\
&= a \left(\theta^2 - \frac{2\theta b}{a} + \frac{b^2}{a^2} \right) - a \left(\frac{b^2}{a^2} \right) + c \\
&= a \left(\theta - \frac{b}{a} \right)^2 - \frac{b^2}{a} + c.
\end{aligned}$$

Therefore

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto e^{\frac{-M}{2}} \\ &= e^{-\frac{1}{2}[a(\theta-\frac{b}{a})^2-\frac{b^2}{a}+c]} \\ &= e^{-\frac{1}{2}[a(\theta-\frac{b}{a})^2]}e^{-\frac{1}{2}[\frac{-b^2}{a}+c]} \\ &= e^{-\frac{1}{2}\frac{(\theta-\frac{b}{a})^2}{\frac{1}{a}}} \underbrace{e^{-\frac{1}{2}[\frac{-b^2}{a}+c]}}_{does\ not\ involve\ \theta}. \end{aligned}$$

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

The posterior density becomes:

$$f(\theta|\mathbf{x}) \propto e^{-\frac{1}{2}\frac{(\theta-\frac{b}{a})^2}{\frac{1}{a}}}.$$

Clearly this is the density of a normal distribution with mean $\frac{b}{a}$ and variance $\frac{1}{a}$.

Therefore

$$\begin{aligned} f(\theta|\mathbf{x}) &= N\left(\ \frac{b}{a}, \frac{1}{a}\ \right) \\ &= N\left(\ \frac{\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\ \right). \end{aligned}$$

Since $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, then $\sum_{i=1}^n x_i = n\bar{x}$.

Therefore

$$f(\theta|\mathbf{x}) = N\left(\ \frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\ \right).$$

Since the normal distribution is symmetric, then the Bayes estimator under absolute loss function is the mean of the posterior distribution.

Hence

$$\hat{\theta} = \frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}.$$

Chapter 2

Priors, Predictions, and Model Choice

2.1 Prior Distributions

In this section, we present some kinds of the prior distributions. Then we will give some examples.

2.1.1 Conjugate Priors

In the Bayesian probability theory, if the posterior distribution $f(\theta|\mathbf{x})$ is in the same family as the prior distribution $f(\theta)$, but may be with different parameters, the prior and posterior are then called ***Conjugate Distributions***. In this case, the prior $f(\theta)$ is called conjugate prior for θ. See [24].

The next Table 2.1 provides some conjugate priors, these cases of the next table help us to find the posterior distribution without calculations.

Table 2.1: Conjugate priors

$f(x_i\|\theta)$ i=1, 2,...,n	prior $f(\theta)$	posterior $f(\theta\|\mathbf{x})$
Normal $N(\theta,\sigma^2)$	Normal $N(\mu,\tau^2)$	Normal $N\left(\frac{\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)$
Poisson $Pn(\theta)$	Gamma $Ga(\alpha,\beta)$	Gamma $Ga(\Sigma_{i=1}^n x_i+\alpha, n+\beta)$
Binomial $Bi(n,\theta)$	Beta $Be(\alpha,\beta)$	Beta $Be(\Sigma_{i=1}^n x_i+\alpha, n^2+\beta-\Sigma_{i=1}^n x_i)$
Exponential $Ex(\theta)$	Gamma $Ga(\alpha,\beta)$	Gamma $Ga(n+\alpha, \Sigma_{i=1}^n x_i+\beta)$

The next example proves the second case of Table 2.1.

Example 2.1.1. [20] *Suppose $X_1, X_2, ..., X_n$ are i.i.d. $Pn(\theta)$, and suppose the prior density for θ 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:

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

$$
\begin{aligned}
&= \frac{1}{x_1!x_2!...x_n!}e^{-n\theta}\theta^{x_1+x_2+...+x_n} \\
&= \frac{1}{x_1!x_2!...x_n!}e^{-n\theta}\theta^{\Sigma_{i=1}^n x_i}.
\end{aligned}
$$

The posterior density is:

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

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

The posterior density becomes:

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

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

$$
f(\theta|\mathbf{x}) = Ga\left(\ \Sigma_{i=1}^n x_i + \alpha, n + \beta\ \right).
$$

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.

2.1.2 Locally Uniform Priors

An improper prior distribution $f(\theta)$ is non-negative for all θ but

$$
\int_{-\infty}^{\infty} f(\theta)d\theta = \infty.
$$

The most common improper priors are

$$f_1(\eta) = C,\ -\infty < \eta < \infty,\ C > 0,$$
$$f_2(\gamma) = \tfrac{1}{\gamma},\ 0 < \gamma < \infty.$$

The improper prior distributions have the following condition

$$\int_{-\infty}^{\infty} f_i(\theta)d\theta \neq 1,\ i = 1, 2.$$

The improper prior distribution is a good choice if the resulting posterior distribution is proper,

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

Further, suppose that

$$f(\theta) = \begin{cases} k, & \text{for values of } \theta; \\ 0, & \text{otherwise.} \end{cases}$$

This $f(\theta)$ then define a proper density. The prior distributions like the above mentioned are called ***locally uniform priors***. See [15].

2.1.3 Non-Informative Priors

If a prior distribution $f(\theta)$ does not contain any information about θ, it is called a **non-informative prior**. Most widely used non-informative priors are ***Jeffreys priors***:

$$f(\theta) = \sqrt{I(\theta)}, \tag{2.1}$$

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

$$I(\theta) = -E_\theta\Big(\tfrac{\partial^2}{\partial\theta^2} \log f(\mathbf{x}|\theta) \Big).$$

See [15].

Example 2.1.2. [9] *Suppose* $X \sim Bi(n, \theta)$.

The likelihood density is:

$$f(x|\theta) \;=\; \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x},\; x \in \{0,1,...,n\},\; n \in \{1,2,...\},\; \theta \in (0,1).$$

Take the logarithm of both sides of the previous equation:

$$\begin{aligned} \log f(x|\theta) &= \log \left(\left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x} \right) \\ &= \log \left(\begin{array}{c} n \\ x \end{array} \right) + \log \theta^x + \log(1-\theta)^{n-x} \\ &= \log \left(\begin{array}{c} n \\ x \end{array} \right) + x \log \theta + (n-x)\log(1-\theta). \end{aligned}$$

We find the first derivative of the previous equation:

$$\frac{\partial}{\partial \theta} \log f(x|\theta) \;=\; \frac{x}{\theta} - \frac{(n-x)}{(1-\theta)},$$

and the second derivative is given by:

$$\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \;=\; -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}.$$

The Fisher information is:

$$\begin{aligned} I(\theta) &= -E_\theta \Big(\; \tfrac{\partial^2}{\partial \theta^2} \log f(X|\theta) \; \Big) \\ &= -E_\theta \Big(\; -\tfrac{X}{\theta^2} - \tfrac{n-X}{(1-\theta)^2} \; \Big) \\ &= -E_\theta \Big(\; -\tfrac{X}{\theta^2} \; \Big) - E_\theta \Big(\; -\tfrac{n-X}{(1-\theta)^2} \; \Big) \\ &= \frac{1}{\theta^2} E_\theta(X) + \frac{1}{(1-\theta)^2} E_\theta(n-X) \\ &= \frac{1}{\theta^2} E_\theta(X) + \frac{1}{(1-\theta)^2} (n - E_\theta(X)) \\ &= \frac{1}{\theta^2} (n\theta) + \frac{1}{(1-\theta)^2} (n - n\theta) \\ &= \frac{n}{\theta} + \frac{n(1-\theta)}{(1-\theta)^2} \\ &= \frac{n}{\theta} + \frac{n}{1-\theta} \\ &= n \Big(\; \tfrac{1}{\theta} + \tfrac{1}{1-\theta} \; \Big). \end{aligned}$$

The Jeffreys prior is:

$$\begin{aligned} f(\theta) &= \sqrt{I(\theta)} \\ &= \sqrt{n\left(\ \tfrac{1}{\theta}+\tfrac{1}{1-\theta}\ \right)} \\ &= \sqrt{n\frac{1}{\theta(1-\theta)}} \\ &= (\sqrt{n})\sqrt{\frac{1}{\theta(1-\theta)}} \\ &\propto \sqrt{\frac{1}{\theta(1-\theta)}} \\ &= \frac{1}{\sqrt{\theta(1-\theta)}}. \end{aligned}$$

Hence

$$f(\theta) \propto \tfrac{1}{\sqrt{\theta(1-\theta)}}.$$

Example 2.1.3. [9] *Suppose $X_1, X_2, ..., X_n$ be i.i.d. $Pn(\theta)$. The likelihood density is:*

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

We take the logarithm of both sides of the previous equation:

$$\begin{aligned}
\log f(\mathbf{x}|\theta) &= \log \frac{\theta^{\Sigma_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} \\
&= \log \theta^{\Sigma_{i=1}^n x_i} + \log e^{-n\theta} - \log \prod_{i=1}^n x_i! \\
&= \Sigma_{i=1}^n x_i \log \theta - n\theta - \log \prod_{i=1}^n x_i!.
\end{aligned}$$

We find the first derivative of the previous equation:

$$\frac{\partial}{\partial \theta} \log f(\mathbf{x}|\theta) = \frac{\Sigma_{i=1}^n x_i}{\theta} - n,$$

and the second derivative is given by:

$$\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{x}|\theta) = -\frac{\Sigma_{i=1}^n x_i}{\theta^2}.$$

The Fisher information is:

$$\begin{aligned}
I(\theta) &= -E_\theta\Big(\tfrac{\partial^2}{\partial \theta^2} \log f(\boldsymbol{X}|\theta) \Big) \\
&= -E_\theta\Big(-\tfrac{\Sigma_{i=1}^n X_i}{\theta^2} \Big) \\
&= -E_\theta \Big(-\tfrac{X_1}{\theta^2} - \tfrac{X_2}{\theta^2} - ..... - \tfrac{X_n}{\theta^2} \Big) \\
&= E_\theta \Big(\tfrac{X_1}{\theta^2} \Big) + E_\theta \Big(\tfrac{X_2}{\theta^2} \Big) + ... + E_\theta \Big(\tfrac{X_n}{\theta^2} \Big) \\
&= \frac{1}{\theta^2} \underbrace{(\theta + \theta + ... + \theta)}_{n\text{-}copies} \\
&= \frac{1}{\theta^2}(n\theta) \\
&= \frac{n}{\theta}.
\end{aligned}$$

The Jeffreys prior is:

$$\begin{aligned}
f(\theta) &\propto \sqrt{I(\theta)} \\
&= \sqrt{\frac{n}{\theta}} \\
&\propto \frac{\sqrt{n}}{\sqrt{\theta}} \\
&= \frac{1}{\sqrt{\theta}}.
\end{aligned}$$

Therfore

$$f(\theta) \propto \frac{1}{\sqrt{\theta}}.$$

2.2 Predictive Distributions

2.2.1 Posterior Predictive Distributions

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

Definition 2.2.1. [15] Let $X_1, X_2, ..., X_n$ be an i.i.d. sample from the distribution $f(x|\theta)$. Let $f(\theta)$ be the prior distribution and $f(\theta|\mathbf{x})$ be the posterior distribution. The posterior predictive distribution is given by:

$$f(x_{n+1}|\mathbf{x}) = \int_{-\infty}^{\infty} f(x_{n+1}|\theta) f(\theta|\mathbf{x}) d\theta. \tag{2.2}$$

Example 2.2.1. [21] *Suppose*

$$\begin{aligned} X_1, X_2, ...X_n &\sim Pn(\theta) \\ \theta &\sim Ga(\alpha, \beta). \end{aligned}$$

Let X_i be independent for all $i = 1, 2, ..., n$ and $\alpha \in \mathbb{Z}^+$.

The likelihood density is:

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

Here

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

Accorrding to Table (2.1)*, the posterior density is:*

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

Hence

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

The posterior predictive density is:

$$\begin{aligned} f(x_{n+1}|\mathbf{x}) &= \int_{-\infty}^{\infty} f(x_{n+1}|\theta)f(\theta|\mathbf{x})d\theta \\ &= \int_{0}^{\infty} \frac{\theta^{x_{n+1}}e^{-\theta}}{x_{n+1}!}\frac{(n+\beta)^{\Sigma_{i=1}^{n}x_i+\alpha}}{\Gamma(\Sigma_{i=1}^{n}x_i+\alpha)}\theta^{\Sigma_{i=1}^{n}x_i+\alpha-1}e^{-(n+\beta)\theta}d\theta \\ &= \frac{(n+\beta)^{\Sigma_{i=1}^{n}x_i+\alpha}}{\Gamma(\Sigma_{i=1}^{n}x_i+\alpha)x_{n+1}!}\int_{0}^{\infty}\theta^{\Sigma_{i=1}^{n+1}x_i+\alpha-1}e^{-(n+\beta+1)\theta}d\theta. \end{aligned}$$

Now this integrand looks like a gamma distribution with parameters:

$\Sigma_{i=1}^{n+1}x_i+\alpha$ *and* $n+\beta+1$,

and since

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

then

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

Therefore

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

Since

$$\begin{aligned} \Gamma(\Sigma_{i=1}^{n+1}x_i+\alpha) &= (\Sigma_{i=1}^{n+1}x_i+\alpha-1)!, \\ \Gamma(\Sigma_{i=1}^{n}x_i+\alpha) &= (\Sigma_{i=1}^{n}x_i+\alpha-1)!, \end{aligned}$$

then

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

And since

$$\frac{(\Sigma_{i=1}^{n+1} x_i+\alpha-1)!}{(\Sigma_{i=1}^{n} x_i+\alpha-1)!(x_{n+1})!} = \begin{pmatrix} \Sigma_{i=1}^{n+1} x_i + \alpha - 1 \\ x_{n+1} \end{pmatrix},$$

then

$$\begin{aligned} f(x_{n+1}|\mathbf{x}) &= \frac{(\Sigma_{i=1}^{n+1} x_i + \alpha - 1)!}{(\Sigma_{i=1}^{n} x_i + \alpha - 1)!(x_{n+1})!} \left(\frac{n+\beta}{n+\beta+1} \right)^{\Sigma_{i=1}^{n} x_i+\alpha} \left(\frac{1}{n+\beta+1} \right)^{x_{n+1}} \\ &= \begin{pmatrix} \Sigma_{i=1}^{n+1} x_i + \alpha - 1 \\ x_{n+1} \end{pmatrix} \left(\frac{n+\beta}{n+\beta+1} \right)^{\Sigma_{i=1}^{n} x_i+\alpha} \left(\frac{1}{n+\beta+1} \right)^{x_{n+1}}. \end{aligned}$$

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

$$f(x_{n+1}|\mathbf{x}) = Nb\left(\Sigma_{i=1}^{n} x_i + \alpha, \frac{n+\beta}{n+\beta+1} \right).$$

Example 2.2.2. [9] *Suppose* $X_1, X_2, ..., X_n$ *be i.i.d.* $N(0, \sigma^2)$ *and the prior for* σ^2 *is given by:*

$$f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\sigma^2)^{\alpha+1}} e^{-\frac{\beta}{\sigma^2}}, \quad \alpha, \beta, \sigma^2 > 0.$$

This is the density of the inverse gamma distribution with parameters: α *and* β.

The likelihood density is:

$$\begin{aligned} f(\mathbf{x}|\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} x_i^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}(x_1^2+...+x_n^2)} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2}. \end{aligned}$$

Therefore the posterior density is:

$$\begin{aligned} f(\sigma^2|\mathbf{x}) &\propto f(\mathbf{x}|\sigma^2) f(\sigma^2) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\sigma^2)^{\alpha+1}} e^{-\frac{\beta}{\sigma^2}} \\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+\alpha+1}} e^{-\frac{1}{\sigma^2}(\beta+\frac{1}{2}\sum_{i=1}^{n} x_i^2)}. \end{aligned}$$

Clearly this is the density of the inverse gamma distribution with parameters $\hat{\alpha} = \frac{n}{2} + \alpha$ *and* $\hat{\beta} = \beta + \frac{1}{2}\sum_{i=1}^{n} x_i^2$.

That is

$$f(x_{n+1}|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x_{n+1}^2}.$$

Note that the posterior distribution $f(\sigma^2|\mathbf{x})$ *is in the same family as the prior distribution* $f(\sigma^2)$ *with different parameters. Therefore* $f(\sigma^2)$ *is conjugate prior for* σ^2.

Then the posterior predictive density is:

$$\begin{aligned} f(x_{n+1}|\mathbf{x}) &= \int_{-\infty}^{\infty} f(x_{n+1}|\sigma^2) f(\sigma^2|\mathbf{x}) d\sigma^2 \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} e^{-\frac{1}{2\sigma^2}x_{n+1}^2} \frac{(\hat{\beta})^{\hat{\alpha}}}{\Gamma(\hat{\alpha})} \frac{1}{(\sigma^2)^{\hat{\alpha}+1}} e^{-\frac{\hat{\beta}}{\sigma^2}} d\sigma^2 \\ &\propto \int_0^{\infty} \frac{1}{(\sigma^2)^{\hat{\alpha}+\frac{1}{2}+1}} e^{-\frac{1}{\sigma^2}(\hat{\beta}+\frac{1}{2}x_{n+1}^2)} d\sigma^2. \end{aligned}$$

Now this integrand looks like the inverse gamma distribution with parameters $\hat{\alpha} + \frac{1}{2}$ *and* $\hat{\beta} + \frac{x_{n+1}^2}{2}$.

Since

$$\int_0^{\infty} \frac{(\hat{\beta} + \frac{1}{2}x_{n+1}^2)^{\hat{\alpha}+\frac{1}{2}}}{\Gamma(\hat{\alpha}+\frac{1}{2})} \frac{1}{(\sigma^2)^{\hat{\alpha}+\frac{1}{2}+1}} e^{-\frac{1}{\sigma^2}(\hat{\beta}+\frac{1}{2}x_{n+1}^2)} d\sigma^2 = 1,$$

then

$$\int_0^{\infty} \frac{1}{(\sigma^2)^{\hat{\alpha}+\frac{1}{2}+1}} e^{-\frac{1}{\sigma^2}(\hat{\beta}+\frac{1}{2}x_{n+1}^2)} d\sigma^2 = \frac{\Gamma(\hat{\alpha}+\frac{1}{2})}{(\hat{\beta}+\frac{1}{2}x_{n+1}^2)^{\hat{\alpha}+\frac{1}{2}}}.$$

Therefore

$$\begin{aligned} f(x_{n+1}|\mathbf{x}) &\propto \frac{\Gamma(\hat{\alpha}+\frac{1}{2})}{(\hat{\beta}+\frac{1}{2}x_{n+1}^2)^{\hat{\alpha}+\frac{1}{2}}} \\ &\propto \frac{1}{\left(\ \hat{\beta}+\frac{1}{2}x_{n+1}^2\ \right)^{\hat{\alpha}+\frac{1}{2}}} \\ &= \left(\ \hat{\beta}+\frac{1}{2}x_{n+1}^2\ \right)^{-\hat{\alpha}-\frac{1}{2}} \\ &= \left(\ (\beta+\frac{1}{2}\sum_{i=1}^{n} x_i^2) + \frac{1}{2}x_{n+1}^2\ \right)^{-(\frac{n}{2}+\alpha)-\frac{1}{2}} \\ &= \left(\ \beta+\frac{1}{2}\sum_{i=1}^{n+1} x_i^2\ \right)^{-\frac{2\alpha+n+1}{2}}. \end{aligned}$$

2.2.2 Prior Predictive Distributions

In this subsection, we present the definition of the prior predictive distribution and give some examples.

Definition 2.2.2. [9] Let $X_1, X_2, ..., X_n$ be an i.i.d. sample from the distribution $f(x|\theta)$. Let $f(\theta)$ be the prior distribution and $f(\theta|\mathbf{x})$ be the posterior distribution. The (joint) prior predictive distribution is given by:

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta) f(\theta) d\theta. \tag{2.3}$$

Note that it is simply the normalizing constant in $f(\theta|\mathbf{x})$.

Example 2.2.3. [20] *Suppose let $X_1, X_2, ..., X_n$ be an i.i.d. sample from the distribution $f(x_i|\theta) = \theta(1-\theta)^{x_i}$, and prior distribution for θ is given by:*

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

The likelihood density is:

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

The (joint) prior predictive density is:

$$\begin{aligned} f(\mathbf{x}) &= \int_0^1 \theta^n(1-\theta)^{\Sigma_{i=1}^n x_i} \frac{1}{beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ &= \frac{1}{beta(\alpha,\beta)} \int_0^1 \theta^{n+\alpha-1}(1-\theta)^{\Sigma_{i=1}^n x_i+\beta-1}d\theta. \end{aligned}$$

Now this integrand looks like a beta distribution with parameters:
$n+\alpha$ and $\Sigma_{i=1}^n x_i + \beta$,

and since

$$\int_0^1 \frac{1}{beta(n+\alpha, \Sigma_{i=1}^n x_i + \beta)} \theta^{n+\alpha-1}(1-\theta)^{\Sigma_{i=1}^n x_i + \beta - 1} d\theta \;\; = \;\; 1,$$

then

$$\int_0^1 \theta^{n+\alpha-1}(1-\theta)^{\Sigma_{i=1}^n x_i + \beta - 1} d\theta \;\; = \;\; beta(n+\alpha, \Sigma_{i=1}^n x_i + \beta),$$

which implies

$$f(\mathbf{x}) \;\; = \;\; \frac{beta(n+\alpha, \Sigma_{i=1}^n x_i + \beta)}{beta(\alpha, \beta)}.$$

Example 2.2.4. *Suppose $X_1, X_2, ..., X_n$ be i.i.d. $N(0, \sigma^2)$, and the prior for σ^2 is given by:*

$$f(\sigma^2) \;\; = \;\; \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\sigma^2)^{\alpha+1}} e^{-\frac{\beta}{\sigma^2}}, \quad \alpha, \beta, \sigma^2 > 0.$$

This is the density of the inverse gamma distribution with parameters: α and β.

The likelihood density is:

$$f(\mathbf{x}|\sigma^2) \;\; = \;\; \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2}, \quad x, \sigma^2 > 0.$$

The (joint) prior predictive density is:

$$\begin{aligned} f(\mathbf{x}) &= \int_{-\infty}^{\infty} f(\mathbf{x}|\sigma^2) f(\sigma^2) d\sigma^2 \\ &= \int_0^{\infty} \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\sigma^2)^{\alpha+1}} e^{-\frac{\beta}{\sigma^2}} d\sigma^2 \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)(2\pi)^{\frac{n}{2}}} \int_0^{\infty} \frac{1}{(\sigma^2)^{\frac{n}{2}+\alpha+1}} e^{-\frac{1}{\sigma^2}(\beta + \frac{1}{2}\sum_{i=1}^n x_i^2)} d\sigma^2. \end{aligned}$$

Now the last integrand looks like the inverse gamma density with parameters $\frac{n}{2} + \alpha$ and $\beta + \frac{1}{2}\sum_{i=1}^n x_i^2$,

and since

$$\int_0^{\infty} \frac{(\beta + \frac{1}{2}\sum_{i=1}^n x_i^2)^{\frac{n}{2}+\alpha}}{\Gamma(\frac{n}{2}+\alpha)} \frac{1}{(\sigma^2)^{\frac{n}{2}+\alpha+1}} e^{-\frac{1}{\sigma^2}(\beta + \frac{1}{2}\sum_{i=1}^n x_i^2)} d\sigma^2 \;\; = \;\; 1,$$

then

$$\int_0^\infty \frac{1}{(\sigma^2)^{\frac{n}{2}+\alpha+1}} e^{-\frac{1}{\sigma^2}(\beta+\frac{1}{2}\sum_{i=1}^n x_i^2)} d\sigma^2 \quad = \quad \frac{\Gamma(\frac{n}{2}+\alpha)}{(\beta+\frac{1}{2}\sum_{i=1}^n x_i^2)^{\frac{n}{2}+\alpha}}.$$

Therefore

$$f(\mathbf{x}) = \frac{\beta^\alpha}{\Gamma(\alpha)(2\pi)^{\frac{n}{2}}} \frac{\Gamma(\frac{n}{2}+\alpha)}{(\beta+\frac{1}{2}\sum_{i=1}^n x_i^2)^{\frac{n}{2}+\alpha}}.$$

2.3 Model Choice

This section consists of two subsections. Firstlly we introduce the definition of the Bayes factor and its use in the choice between two models. Secondly we introduce the use of Bayes factor in the choice between two hypotheses.

2.3.1 Bayes Factors

Suppose that we have to choose between two Models M_0 and M_1 for data, and let $\mathbf{x} = (x_1, x_2, ..., x_n)$.
The likelihoods are denoted by:

$$f_i(\mathbf{x}|\theta_i),\ i = 0, 1,$$

and the priors by:

$$f_i(\theta_i),\ i = 0, 1.$$

Recall that the prior predictive distribution for model i is:

$$f(\mathbf{x}|M_i) = \int_{-\infty}^{\infty} f_i(\mathbf{x}|\theta_i) f_i(\theta_i) d\theta_i.$$

The Bayes factor is defined as:

$$B_{01} = \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}. \tag{2.4}$$

Bayes factor has three cases:

1. If $B_{01} > 1$, accept M_0.

2. If $B_{01} < 1$, accept M_1.

3. If $B_{01} = 1$, undecidability.

Example 2.3.1 (Geometric versus Poisson)**.** [9] *Suppose that*

$M_0 : X_1, X_2, ..., X_n|\theta_0 \sim f_0(x|\theta_0) = \theta_0(1-\theta_0)^x$, *x = 0, 1,...*

$M_1 : X_1, X_2, ..., X_n|\theta_1 \sim f_1(x|\theta_1) = \frac{e^{-\theta_1}{\theta_1}^x}{x!}$, *x = 0, 1,...*

Further, assume that θ_0 *and* θ_1 *are unknown.*
How should we decide between the two models based on $x_1, x_2, ..., x_n$*? The likelihood density for model* M_0

$$\begin{aligned} f_0(\mathbf{x}|\theta_0) &= \prod_{i=1}^{n} f_0(x_i|\theta_0) \\ &= \prod_{i=1}^{n} \theta_0(1-\theta_0)^{x_i} \\ &= \underbrace{\theta_0\theta_0...\theta_0}_{n\text{-}copies}(1-\theta_0)^{x_1}(1-\theta_0)^{x_2}...(1-\theta_0)^{x_n} \\ &= {\theta_0}^n(1-\theta_0)^{\sum_{i=1}^n x_i}. \end{aligned}$$

Assume that $f_0(\theta_0)$ *is the beta distribution with parameters* α_0 *and* β_0*. And* $f_1(\theta_1)$ *is the gamma distribution with parameters* α_1 *and* β_1*.*
The prior predictive distribution for model M_0

$$\begin{aligned} f(\mathbf{x}|M_0) &= \int_{-\infty}^{\infty} f_0(\mathbf{x}|\theta_0)f_0(\theta_0)d\theta_0 \\ &= \int_0^1 {\theta_0}^n(1-\theta_0)^{\sum_{i=1}^n x_i}\frac{1}{beta(\alpha_0,\beta_0)}{\theta_0}^{\alpha_0-1}(1-\theta_0)^{\beta_0-1}d\theta_0 \\ &= \frac{1}{beta(\alpha_0,\beta_0)}\int_0^1 {\theta_0}^{n+\alpha_0-1}(1-\theta_0)^{\sum_{i=1}^n x_i+\beta_0-1}d\theta_0. \end{aligned}$$

Now this integrand looks like the beta distribution with parameters:
$n + \alpha_0$ *and* $\Sigma_{i=1}^n x_i + \beta_0$,

and since

$$\int_0^1 \frac{1}{beta(n+\alpha_0, \Sigma_{i=1}^n x_i + \beta_0)} {\theta_0}^{n+\alpha_0-1}(1-\theta_0)^{\Sigma_{i=1}^n x_i+\beta_0-1} d\theta_0 \quad = \quad 1,$$

then

$$\int_0^1 {\theta_0}^{n+\alpha_0-1}(1-\theta_0)^{\Sigma_{i=1}^n x_i+\beta_0-1} d\theta_0 \quad = \quad beta(n+\alpha_0, \Sigma_{i=1}^n x_i + \beta_0).$$

Therefore

$$f(\mathbf{x}|M_0) \quad = \quad \frac{beta(n+\alpha_0, \Sigma_{i=1}^n x_i + \beta_0)}{beta(\alpha_0, \beta_0)}.$$

The likelihood density for model M_1

$$\begin{aligned} f_1(\mathbf{x}|\theta_1) &= \prod_{i=1}^n f_1(x_i|\theta_1) \\ &= \prod_{i=1}^n \frac{e^{-\theta_1}{\theta_1}^{x_i}}{x_i!} \\ &= \underbrace{e^{-\theta_1}e^{-\theta_1}...e^{-\theta_1}}_{n\text{-}copies} {\theta_1}^{x_1}{\theta_1}^{x_2}...{\theta_1}^{x_n}\frac{1}{x_1!x_2!...x_n!} \\ &= \frac{1}{\prod_{i=1}^n x_i!} e^{-n\theta_1}{\theta_1}^{\Sigma_{i=1}^n x_i}. \end{aligned}$$

The prior predictive distribution for model M_1

$$\begin{aligned} f(\mathbf{x}|M_1) &= \int_{-\infty}^{\infty} f_1(\mathbf{x}|\theta_1) f_1(\theta_1) d\theta_1 \\ &= \int_0^{\infty} \frac{{\theta_1}^{\Sigma_{i=1}^n x_i} e^{-n\theta_1}}{\Pi_{i=1}^n (x_i!)} \frac{{\beta_1}^{\alpha_1}}{\Gamma(\alpha_1)} {\theta_1}^{\alpha_1-1} e^{-\beta_1\theta_1} d\theta_1 \\ &= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)\Pi_{i=1}^n x_i!} \int_0^{\infty} {\theta_1}^{\Sigma_{i=1}^n x_i+\alpha_1-1} e^{-(n+\beta_1)\theta_1} d\theta_1, \end{aligned}$$

Now this integrand looks like a gamma distribution with parameters:

$\Sigma_{i=1}^n x_i + \alpha_1$ *and* $n + \beta_1$,

and since

$$\int_0^{\infty} \frac{(n+\beta_1)^{\Sigma_{i=1}^n x_i+\alpha_1}}{\Gamma(\Sigma_{i=1}^n x_i + \alpha_1)} {\theta_1}^{\Sigma_{i=1}^n x_i+\alpha_1-1} e^{-(n+\beta_1)\theta_1} d\theta_1 \quad = \quad 1,$$

which implies

$$\int_0^\infty {\theta_1}^{\Sigma_{i=1}^n x_i+\alpha_1-1} e^{-(n+\beta_1)\theta_1} d\theta_1 \;=\; \frac{\Gamma(\Sigma_{i=1}^n x_i+\alpha_1)}{(n+\beta_1)^{\Sigma_{i=1}^n x_i+\alpha_1}}.$$

Therefore

$$f(\mathbf{x}|M_1) \;=\; \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)\Pi_{i=1}^n x_i!}\frac{\Gamma(\Sigma_{i=1}^n x_i+\alpha_1)}{(n+\beta_1)^{\Sigma_{i=1}^n x_i+\alpha_1}}.$$

Then the Bayes factor is:

$$\begin{aligned} B_{01} &= \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)} \\ &= \frac{beta(n+\alpha_0, \Sigma_{i=1}^n x_i+\beta_0)\Gamma(\alpha_1)(n+\beta_1)^{\Sigma_{i=1}^n x_i+\alpha_1}\Pi_{i=1}^n x_i!}{beta(\alpha_0,\beta_0)\beta_1^{\alpha_1}\Gamma(\Sigma_{i=1}^n x_i+\alpha_1)}. \end{aligned}$$

We now calculate numerical values of Bayes factor.

Let $n=2$, $x_1=2, x_2=2$, $\alpha_0=1$, $\alpha_1=2$, $\beta_0=2$, *and* $\beta_1=1$.

Therefore

$$\begin{aligned} B_{01}(\mathbf{x}) &= \frac{beta(2+1,4+2)\Gamma(2)(2+1)^{(4+2)}(2!2!)}{beta(1,2)1^2\Gamma(4+2)} \\ &= \frac{beta(3,6)(3)^6(4)}{beta(1,2)(5!)} \\ &= \frac{(0.006)(729)(4)}{(0.5)(120)}. \\ &= 0.2916. \end{aligned}$$

Therefore we accept M_1. □

Let $P(M_i)$ denote the prior probability for model $i=0,1$. Let us now calculate the posterior probability of M_i given the data using the Bayes theorem.

$$P(M_i|\mathbf{x}) \;=\; \frac{P(M_i)f(\mathbf{x}|M_i)}{\sum_{j=0}^1 P(M_j)f(\mathbf{x}|M_j)}.$$

So the posterior odds ratio of the two models is given by

$$\frac{P(M_0|\mathbf{x})}{P(M_1|\mathbf{x})} \;=\; \frac{P(M_0)}{P(M_1)} \times \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}.$$

In the words,

$$\text{posterior odds ratio} = \text{prior odds ratio} \times \text{Bayes factor}.$$

We can define the Bayes factor as

$$Bayes factor = \frac{posterior\ odds\ ratio}{prior\ odds\ ratio}.$$

2.3.2 Hypothesis Testing

Let us begin with the definition of a statistical hypothesis.

Definition 2.3.1. [13] A hypothesis is a statement about a population parameter.

Definition 2.3.2. [13] The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

Suppose that we wish to test

$$H_0 : \theta_0 \in \Theta_0 \text{ versus } H_1 : \theta_1 \in \Theta_1.$$

Let $f(\mathbf{x}|\theta)$ denoted the likelihood of $\mathbf{x}$ given θ.
Bayes factor has the following cases:

1. $B_{01}(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)}$ (simple versus simple test).
2. $B_{01}(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_0)}{\int_{\Theta_1} f(\mathbf{x}|\theta_1)f(\theta_1)d\theta_1}$ (simple versus composite test).
3. $B_{01}(\mathbf{x}) = \frac{\int_{\Theta_0} f(\mathbf{x}|\theta_0)f(\theta_0)d\theta_0}{\int_{\Theta_1} f(\mathbf{x}|\theta_1)f(\theta_1)d\theta_1}$ (composite versus composite test).

Bayes factor has three cases:

1. If $B_{01} > 1$, accept M_0.
2. If $B_{01} < 1$, accept M_1.
3. If $B_{01} = 1$, undecidability.

Example 2.3.2. [9] *Let $X_1, X_2, ..., X_6$ be a sequence of independent identically distributed Bernoulli random variables with parameter θ, and suppose that $x_1 = x_2 = x_3 = x_4 = x_5 = 1$ and $x_6 = 0$. Assume that the prior distribution is given by*

$$f(\theta_1) = 2,\ \theta_1 \in (\tfrac{1}{2}, 1).$$

Our problem is

$$H_0 : \theta_0 = \tfrac{1}{2} \text{ versus } H_1 : \theta_1 > \tfrac{1}{2}.$$

It is simple versus composite test.

Since the parameter is known under H_0, we do not need to assume any prior.

The likelihood density under H_0 is

$$\begin{aligned} f(\mathbf{x}|\theta_0) &= \prod_{i=1}^{n} f(x_i|\theta_0) \\ &= \left(\tfrac{1}{2} \right)^6 \\ &= \frac{1}{64}. \end{aligned}$$

The likelihood density under H_1 is

$$\begin{aligned} f(\mathbf{x}|\theta_1) &= \prod_{i=1}^{n} f(x_i|\theta_1) \\ &= {\theta_1}^5(1-\theta_1). \end{aligned}$$

The Bayes factor is

$$\begin{aligned} B_{01}(\mathbf{x}) &= \frac{f(\mathbf{x}|\theta_0)}{\int_{\Theta_1} f(\mathbf{x}|\theta_1) f(\theta_1) d\theta_1} \\ &= \frac{\frac{1}{64}}{\int_{\frac{1}{2}}^{1} 2{\theta_1}^5(1-\theta_1) d\theta_1} \\ &= 0.35. \end{aligned}$$

Therefore there is some evidence in favor of H_1.

Example 2.3.3. [9] *Suppose that $X_1, X_2, ..., X_n$ be an i.i.d. $Nb(x|r,\theta)$ distribution with the pdf*

$$f(x|r,\theta) = \begin{pmatrix} r+x-1 \\ x \end{pmatrix} \theta^r (1-\theta)^x, x = 0, 1, ..., 0 < \theta < 1.$$

where $r > 0$ is a known integer. Suppose also the prior has $Be(\alpha, \beta)$ distribution with the pdf

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

The likelihood density is:

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} \begin{pmatrix} r+x_i-1 \\ x_i \end{pmatrix} \theta^r (1-\theta)^{x_i} \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} r+x_i-1 \\ x_i \end{pmatrix} \right\} \underbrace{\theta^r \theta^r ... \theta^r}_{n-times} (1-\theta)^{x_1} (1-\theta)^{x_2} ... (1-\theta)^{x_n} \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} r+x_i-1 \\ x_i \end{pmatrix} \right\} \theta^{nr} (1-\theta)^{\Sigma_{i=1}^n x_i}. \end{aligned}$$

The posterior density is:

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta) f(\theta) \\ &= \left\{ \prod_{i=1}^{n} \begin{pmatrix} r+x_i-1 \\ x_i \end{pmatrix} \right\} \theta^{nr} (1-\theta)^{\Sigma_{i=1}^n x_i} \frac{1}{beta(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{1}{beta(\alpha,\beta)} \left\{ \prod_{i=1}^{n} \begin{pmatrix} r+x_i-1 \\ x_i \end{pmatrix} \right\} \theta^{nr} (1-\theta)^{\Sigma_{i=1}^n x_i} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \underbrace{\frac{1}{beta(\alpha,\beta)} \left\{ \prod_{i=1}^{n} \begin{pmatrix} r+x_i-1 \\ x_i \end{pmatrix} \right\}}_{does\ not\ involve\ \theta} \theta^{nr+\alpha-1} (1-\theta)^{\Sigma_{i=1}^n x_i+\beta-1}. \end{aligned}$$

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

Therefore the posterior distribution becomes:

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

It is the density of a beta distribution with parameters $nr+\alpha$ *and* $\Sigma_{i=1}^n x_i+\beta$.
That is

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

Suppose further that $r=2, n=1, \alpha=\beta=1$ *and we observe that* $x_1=1$.
Of the two hypotheses $H_0:\theta_0 \leq \frac{1}{2}$ *and* $H_1:\theta_1 > \frac{1}{2}$, *what is the Bayes factor in favor of* H_1?
It is composite versus composite test.
Let us find $f(x_1|H_0)$ *and* $f(x_1|H_1)$.
Here $f(x_1|H_0)$ *is given by:*

$$\begin{aligned} f(x_1|H_0) &= \int_0^{\frac{1}{2}} f(x_1|\theta_0)f(\theta_0)d\theta_0 \\ &= \int_0^{\frac{1}{2}} 2{\theta_0}^2(1-\theta_0)2d\theta_0 \\ &= 4\int_0^{\frac{1}{2}} {\theta_0}^2(1-\theta_0)d\theta_0 = 0.3125. \end{aligned}$$

And $f(x_1|H_1)$ *is*

$$\begin{aligned} f(x_1|H_1) &= \int_{\frac{1}{2}}^1 f(x_1|\theta_1)f(\theta_1)d\theta_1 \\ &= \int_{\frac{1}{2}}^1 2{\theta_1}^2(1-\theta_1)2d\theta_1 \\ &= 4\int_{\frac{1}{2}}^1 {\theta_1}^2(1-\theta_1)d\theta_1 = 0.6875. \end{aligned}$$

The Bayes factor in favor of H_1 *is:*

$$\begin{aligned} B_{01}(x_1) &= \frac{f(x_1|H_0)}{f(x_1|H_1)} \\ &= \frac{0.3125}{0.6875} \\ &= 0.4545. \end{aligned}$$

Therefore there is some evidence in favor of H_1.

Chapter 3

Markov Chains

A Markov chain is a discrete stochastic process with a finite state space and a property called the *Markov Property* as we will formally define below. It is usually characterized as memoryless: The next state depends only on the current state. The importance of Markov chains comes from this fact that they can be used to statistically model real-world processes.

3.1 Definitions and Basic Properties

In this chapter, we are interested in studying discrete Markov chains which have a finite state space. Suppose that the random variables $X_0, X_1, \ldots, X_n$ represent the outcomes of some random experiments. We assume that the outcomes belong to $\Omega = \{x_0, x_1, \ldots, x_n\}$, where Ω is a state space for this system. The result of any experiment depends only on the immediate previous result.
We denote by $p(i, j)$ the probability of moving from the state i to the state j in one step.
More precisely, $p(i, j) = P(X_{n+1} = j | X_n = i)$. Such these stochastic experiments are called a ***finite Markov chains***. See [6].

Definition 3.1.1. [24] Let $\{X_0, X_1, \ldots, X_n\}$ be a stochastic process, a collection of

random variables indexed be times, n. If the stochastic process with a discrete state space $\Omega = \{x_0, x_1, \ldots, x_n\}$ has the property that $\forall n \in \mathbb{N}$ and $\forall x_0, x_1, ..., i, j \in \Omega$,

$$P(X_{n+a} = j | X_0 = x_0, X_1 = x_1, ..., X_n = i) = P(X_{n+a} = j | X_n = i), \forall a > 0. \quad (3.1)$$

Then this process is said to be a ***Markov chain***.

Definition 3.1.2. [24] Consider $\{X_0, X_1, \ldots, X_n\}$ be a Markov chain with a discrete state space $\Omega = \{x_0, x_1, \ldots, x_n\}$, then the matrix $\mathrm{T} = \Big(\ p(i,j) \ \Big)$, where $p(i,j) = P(X_{n+1} = j | X_n = i)$ characterized the transition kernal of a Markov chain, and is reffered to as a transition matrix.

The transition matrix T has following properties:

$$p(i,j) \geq 0, \qquad \forall i, j \in \Omega. \quad (3.2)$$

$$\sum_{j \in \Omega} p(i,\ j) = 1, \qquad \forall i \in \Omega. \quad (3.3)$$

In words, the last two properties say:

1. The entries of the matrix are nonnegative.

2. Each row of the transition matrix sum to 1.

Example 3.1.1 (Gambler's ruin)**.** [19] *Consider a gambling game in which on any turn you win* 1$ *with probability* $p = 0.4$ *or lose* 1$ *with probability* $1 - p = 0.6$. *Suppose further that you adopt the rule that you quit playing if your fortune reaches* $N$$. *Of course, if your fortune reaches* 0$ *the casino makes you stop. Let* X_n *be the amount of money you have after n plays. Your fortune,* X_n *has the Markov property. To check this for the gambler's ruin chain, we note that if you are still playing at time n, i.e., your fortune* $X_n = i$ *with* $0 < i < N$, *then for any possible history your wealth* $i_{n-1}, i_{n-2}, ..., i_1, i_0$

$$P(X_{n+1} = i + 1 | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) \ = \ 0.4.$$

We say that X_n *is a discrete time Markov chain with transition matrix* T *if for any* $j, i, i_{n-1}, i_{n-2}, ..., i_1, i_0$ *we have*

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

The transition probabilities:

$$p(i,j) = \begin{cases} 0.4, & \text{if } j = i+1, \quad 0 < i < N; \\ 0.6, & \text{if } j = i-1, \quad 0 < i < N; \\ 1, & \text{if } i = j = 0; \\ 1, & \text{if } i = j = N \end{cases}$$

If $N = 5$*, then the transition matrix is:*

$$\mathrm{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

which has a transition graph as follows

Figure 3.1: Transition graph for a Gambler's ruin chain

Example 3.1.2 (Ehrenfest chain)**.** [19] *This chain originated in physics as a model for two cubical volumes of air connected by a small hole. In the mathematical version, we have two urns, i.e., two of the exalted trash cans of probability theory, in which there are a total of N balls. We pick one of the N balls at random and move it to the other urn.*

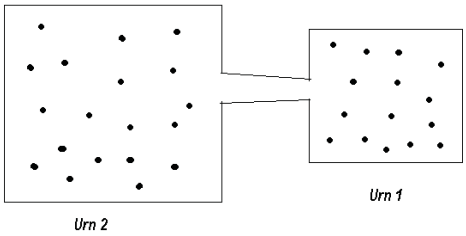

Figure 3.2: Efrenfest chain sketch

Let X_n be the number of balls in the left urn after the nth draw. It should be clear that X_n has the Markov property; i.e., if we want to guess the state at time $n+1$, then the current number of balls in the left urn X_n, is the only relevant information from the observed sequence of states $X_n, X_{n-1}, ..., X_1, X_0$. To check this we note that

$$P(X_{n+1} = i+1 | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) \quad = \quad \frac{N-i}{N}.$$

Since to increase the number we have to pick one of the $N-i$ balls in the other urn. The number can also decrease by 1 with probability $\frac{i}{N}$. In symbols, we have computed that the transition probability is given by:

$$p(i,j) = \begin{cases} \frac{N-i}{N}, & \text{if } j = i+1,\ 0 \le i \le N; \\ \frac{i}{N}, & \text{if } j = i-1,\ 1 \le i \le N; \\ 0, & \text{otherwie.} \end{cases}$$

If $N = 4$, then the transition matrix is

$$\mathrm{T} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 \\ 0 & \frac{2}{4} & 0 & \frac{2}{4} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

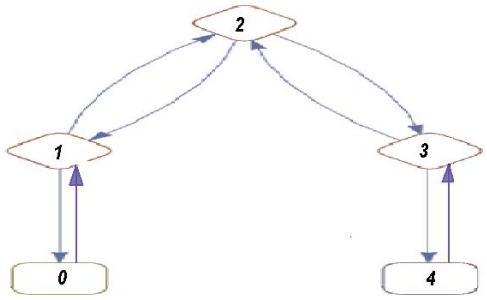

Figure 3.3: Transition graph for a Ehrenfest chain

3.2 Multistep Transition Probabilities

Suppose that $\{X_0, X_1, \ldots, X_n\}$ be a Markov chain with a discrete state space $\Omega = \{x_0, x_1, \ldots, x_n\}$ and $p(i,j) = P(X_{n+1} = j \mid X_n = i)$ gives the probability of going from i to j at one step. In this section we compute the probability of going from i to j at $m > 1$ steps

$$p^m(i,j) = P(X_{n+m} = j \mid X_n = i). \tag{3.4}$$

Theorem 3.2.1 (Chapman-Kolmogorov equation)**.** [23] $\forall i, j \in \Omega = \{x_0, x_1, ...\}$, *we have that*

$$p^{m+n}(i,\, j) = \sum_{k \in \Omega} p^m(i,\, k)\, p^n(k,\, j). \tag{3.5}$$

Proof. To go from i to j in $m+n$ steps, we have to go from i to some state k in m steps and then from k to j in n steps. The Markov property implies that the two parts of our journey are independent.

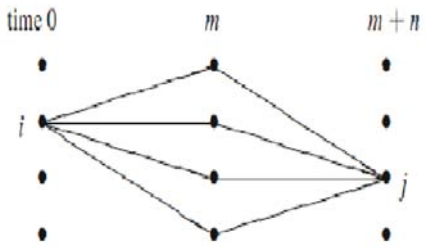

Figure 3.4: Chapman-Kolmogorov

$$
\begin{aligned}
p^{m+n}(i,j) &= P(X_{m+n}=j|X_0=i) \\
&= \sum_{k\in\Omega} P(X_{m+n}=j, X_m=k|X_0=i) \\
&= \sum_{k\in\Omega} \frac{P(X_{m+n}=j, X_m=k, X_0=i)}{P(X_0=i)} \\
&= \sum_{k\in\Omega} \frac{P(X_{m+n}=j, X_m=k, X_0=i)}{P(X_0=i)} \frac{P(X_m=k, X_0=i)}{P(X_m=k, X_0=i)} \\
&= \sum_{k\in\Omega} \frac{P(X_{m+n}=j, X_m=k, X_0=i)}{P(X_m=k, X_0=i)} \frac{P(X_m=k, X_0=i)}{P(X_0=i)} \\
&= \sum_{k\in\Omega} P(X_{m+n}=j|X_m=k, X_0=i)\, P(X_m=k|X_0=i) \\
&= \sum_{k\in\Omega} P(X_{m+n}=j|X_m=k)\, P(X_m=k|X_0=i) \\
&= \sum_{k\in\Omega} p^n(k,\, j)\, p^m(i,\, k).
\end{aligned}
$$

Therefore

$$
p^{m+n}(i,\, j) = \sum_{k\in\Omega} p^m(i,\, k)\, p^n(k,\, j).
$$

Note that $p^{m+n}(i,j)$ is the ij^{th} entry of the matrix T^{m+n}, $p^m(i,k)$ is the ik^{th} entry of the matrix T^m, and $p^n(k,j)$ is the kj^{th} entry of the matrix T^n. As we know, from the definition of the product of two matrices, the defining relation for ij^{th} entry of the product T^m and T^n is identical to (3.6).

Hence the Chapman-Kolmogorov equation, in matrix from, is:

$$
\mathrm{T}^{m+n} = \mathrm{T}^m\, \mathrm{T}^n. \tag{3.6}
$$

□

Definition 3.2.1 (Regular transition matrix)**.** [22] A transition matrix is ***regular*** if

$$\exists\ n \in \mathbb{Z}^+ \text{ such that } \mathrm{T}^n > 0.$$

In the words a transition matrix is regular if some power of T contains only positive entries.

Example 3.2.1 (Weather chain)**.** [19] *Let X_n be the weather on day n, which we assume is either: $1 =$ rainy, $2 =$ sunny. Even though the weather is not exactly a Markov chain model for weather by writing down a transition matrix*

$$\mathrm{T} = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}.$$

T *has the following properties:*

$$p(i, j) \geq 0, \qquad \forall i, j \in \{1, 2\},$$

$$\sum_{j \in \{1,\,2\}} p(1,\, j) = p(1,\, 1) + p(1,\, 2) = 0.6 + 0.4 = 1,$$

and

$$\sum_{j \in \{1,\,2\}} p(2,\, j) = p(2,\, 1) + p(2,\, 2) = 0.2 + 0.8 = 1.$$

We can use the Chapman-Kolmogorov equation to compute T^2

$$\begin{aligned} p^2(1,\,1) &= \sum_{k=1}^{2} p(1,\, k)\, p(k,\, 1) \\ &= p(1,\, 1)\, p(1,\, 1) + p(1,\, 2)\, p(2,\, 1) \\ &= (0.6 \times 0.6) + (0.4 \times 0.2) \\ &= 0.44. \end{aligned}$$

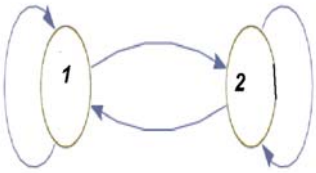

Figure 3.5: Transition graph for a Weather chain

$$\begin{aligned} p^2(1,\,2) &= \sum_{k=1}^{2} p(1,\,k)\,p(k,\,2) \\ &= p(1,\,1)\,p(1,\,2) + p(1,\,2)\,p(2,2) \\ &= (0.6 \times 0.4) + (0.4 \times 0.8) \\ &= 0.56. \end{aligned}$$

$$\begin{aligned} p^2(2,\,1) &= \sum_{k=1}^{2} p(2,\,k)\,p(k,\,1) \\ &= p(2,\,1)\,p(1,\,1) + p(2,\,2)\,p(2,\,1) \\ &= (0.2 \times 0.6) + (0.8 \times 0.2) \\ &= 0.28. \end{aligned}$$

$$\begin{aligned} p^2(2,\,2) &= \sum_{k=1}^{2} p(2,\,k)\,p(k,\,2) \\ &= p(2,\,1)\,p(1,\,2) + p(2,\,2)\,p(2,\,2) \\ &= (0.2 \times 0.4) + (0.8 \times 0.8) \\ &= 0.72. \end{aligned}$$

Therefore

$$\mathrm{T}^2 = \begin{pmatrix} 0.44 & 0.56 \\ 0.28 & 0.72 \end{pmatrix}.$$

Since $\mathrm{T}^2 > 0$*, then* T *is regular transition matrix.*

T^2 *has the following properties:*

$$\forall i, j \in \{1,\, 2\},\, p^2(i, j) \geq 0,$$

$$\sum_{j \in \{1,2\}} p^2(1,\, j) = p^2(1,\, 1) + p^2(1,\, 2) = 0.44 + 0.56 = 1,$$

and

$$\sum_{j \in \{1,2\}} p^2(2,\, j) = p^2(2,\, 1) + p^2(2,\, 2) = 0.28 + 0.72 = 1.$$

$$\mathrm{T}\,\mathrm{T} = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} = \begin{pmatrix} 0.44 & 0.56 \\ 0.28 & 0.72 \end{pmatrix} = \mathrm{T}^2.$$

Let $\{X_0, X_1, \ldots, X_m\}$ be a Markov chain with a transition matrix T. The following theorem shows that if the probability mass function of X_0 is known then, we can find the probability mass function of X_n, $\forall n \geq 1$.

Theorem 3.2.2. [23] *Let $\{X_0, X_1, ..., X_n\}$ be a Markov chain with a transition matrix* T. *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*

$$P(X_n = j) \;=\; \sum_{i \in \Omega} p_i\, p^n(i,\, j), \qquad j = 0, 1, 2, ... \tag{3.7}$$

Proof. Applying the law of total probability Theorem (1.2.2) to the sequence of mutually exclusive events $\{X_0 = i\}$, $i \geq 0$, we have

$$\begin{aligned} P(X_n = j) &= \sum_{i \in \Omega} P(X_n = j | X_0 = i)\, P(X_0 = i) \\ &= \sum_{i \in \Omega} p^n(i,\, j)\, p_i \\ &= \sum_{i \in \Omega} p_i\, p^n(i,\, j) \end{aligned}$$

□

Example 3.2.2. [23] *Suppose that a mouse is moving inside the maze shown in Figure* (3.5)*, from one cell to another, in search of food. When at a cell, the mouse*

will move to one of the adjoining cells randomly. For $n \geq 0$*, let* X_n *be the cell number the mouse will visit after having changed cells* n *times. Then* $\{X_n : n = 0, 1, ...\}$ *is a Markov chain with a state space* $\Omega = \{1, 2, ..., 9\}$ *and a transition matrix*

$$\mathrm{T} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

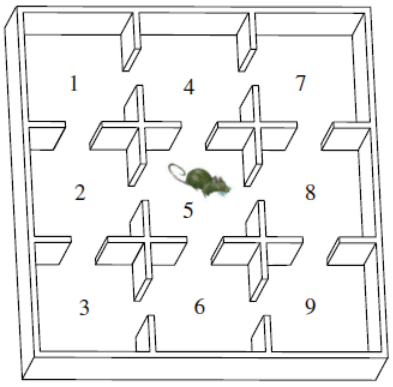

Figure 3.6: The moving mouse of Example 3.2.2

It is equally likely that the mouse is in any of the 9 *cells.*

That is,

$$p_i = P(X_0 = i) = \tfrac{1}{9},\ \textit{for}\ 1 \leq i \leq 9.$$

$$\mathrm{T}^5 = \begin{pmatrix} 0 & \frac{5}{18} & 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 & \frac{2}{9} & 0 \\ \frac{5}{27} & 0 & \frac{5}{27} & 0 & \frac{1}{3} & 0 & \frac{4}{27} & 0 & \frac{4}{27} \\ 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 \\ \frac{5}{27} & 0 & \frac{4}{27} & 0 & \frac{1}{3} & 0 & \frac{5}{27} & 0 & \frac{4}{27} \\ 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ \frac{4}{27} & 0 & \frac{5}{27} & 0 & \frac{1}{3} & 0 & \frac{4}{27} & 0 & \frac{5}{27} \\ 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 \\ \frac{4}{27} & 0 & \frac{4}{27} & 0 & \frac{1}{3} & 0 & \frac{5}{27} & 0 & \frac{5}{27} \\ 0 & \frac{2}{9} & 0 & \frac{2}{9} & 0 & \frac{5}{18} & 0 & \frac{5}{18} & 0 \end{pmatrix}.$$

We can find the probability that the mouse is in cell j, $1 \leq j \leq 9$, *after* 5 *transitions. For example,*

$$\begin{aligned} P(X_5 = 4) &= \sum_{i=1}^{9} p_i p^5(i,4) \\ &= \frac{1}{9} \sum_{i=1}^{9} p^5(i,4) \\ &= \frac{1}{9} \left(\frac{5}{18} + 0 + \frac{2}{9} + 0 + \frac{1}{4} + 0 + \frac{5}{18} + 0 + \frac{2}{9} \right) \\ &= 0.13 \end{aligned}$$

3.3 Classification of States

We say that a state j is ***accessible*** from state i, $i \to j$ if

$$p^n(i,j) > 0 \text{ for some } n \geq 0.$$

This means that there is a possibility of reaching j from i in some number of steps. If j is not accessible from state i, $i \nrightarrow j$, then

$$p^n(i,j) = 0, \, \forall n \geq 0,$$

and thus the chain started from i never visits j.

Theorem 3.3.1. [19] *If $x \to y$ and $y \to z$ then $x \to z$.*

Proof. Let $\{X_0,\, X_1,...\}$ be a Markov chain with a state space $\Omega = \{x_0, x_1, ...\}$ and $x, y, z \in E$.
Since $x \to y$ then $\exists n \geq 1$ such that $p^n(x, y) > 0$,
and since $y \to z$ then $\exists s \geq 1$ such that $p^s(y, z) > 0$,
by the Chapman-Kolmogorov equation,

$$\begin{aligned} p^{n+s}(x,\, z) &= \sum_{m \in \Omega} p^n(x,\, m)\, p^s(m,\, z) \\ &> p^n(x,\, y)\, p^s(y,\, z) > 0. \end{aligned}$$

Therefore $x \to z$. □

Definition 3.3.1. [23] We say that i and j ***coummunicate***, $i \leftrightarrow j$, if

$$\exists m, n > 0 \text{ such that } p^n(i, j) > 0 \text{ and } p^m(j, i) > 0.$$

In words, i and j coummunicate if i is accessible from j, and j is accessible from i.

It is easy to check that this is an equivalence relation:

- *Reflexivity:* Since $p^0(i, i) = 1$ then $\forall i \in \Omega,\ i \leftrightarrow i$.
- *Symmetry:* If $i \leftrightarrow j$ then $\exists m, n > 0$ such that $p^n(i, j) > 0$ and $p^m(j, i) > 0$, therefore $p^m(j, i) > 0$ and $p^n(i, j) > 0$ showing $j \leftrightarrow i$.
- *Transitivity:* Let $\{X_0,\, X_1,...,\, X_n\}$ be a Markov chain with a state space $\Omega = \{x_0, x_1, ..., x_n\}$ and $i, j, k \in \Omega$.
  If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$, to show this,
  we will establish that $i \to k$,
  since $i \leftrightarrow j$ then $\exists n, s > 0$ such that $p^n(i,\, j) > 0$ and $p^s(j,\, i) > 0$,
  since $j \leftrightarrow k$ then $\exists \grave{n}, \grave{s} > 0$ such that $p^{\grave{n}}(j,\, k) > 0$ and $p^{\grave{s}}(k,\, j) > 0$,

by the Chapman-Kolmogorov equation,

$$\begin{aligned} p^{n+\grave{n}}(i,\, k) &= \sum_{l\in\Omega} p^{n}(i,\, l)\, p^{\grave{n}}(l,\, k) \\ &\geq p^{n}(i,\, j)\, p^{\grave{n}}(j,\, k) \\ &> 0, \end{aligned}$$

showing $i \to k$. And

$$\begin{aligned} p^{s+\grave{s}}(k,\, i) &= \sum_{l\in\Omega} p^{\grave{s}}(k,\, l)\, p^{s}(l,\, i) \\ &\geq p^{\grave{s}}(k,\, j)\, p^{s}(j,\, i) \\ &> 0, \end{aligned}$$

showing $k \to i$.

Hence $i \leftrightarrow k$.

Definition 3.3.2. [24] We call $A \subset \Omega$ of states ***closed*** if

$$p(i,j) = 0, \qquad \forall i \in A \, and \, \forall j \notin A. \tag{3.8}$$

Definition 3.3.3. [14] A Markov chain with a state space $\Omega = \{x_0, x_1, ...\}$ and a transition matrix T is an ***irreducible*** if

$$i \leftrightarrow j, \qquad \forall i,\, j \in \Omega. \tag{3.9}$$

Example 3.3.1. *Consider a Markov chain with a state space* $\Omega = \{1, 2, 3\}$ *and a transition matrix*

$$\mathrm{T} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

$p(1,2) = \frac{1}{2} > 0$ *and* $p(2,1) = \frac{1}{2} > 0$ *then* $1 \leftrightarrow 2$,
$p(2,3) = \frac{1}{2} > 0$ *and* $p(3,2) = \frac{1}{3} > 0$ *then* $2 \leftrightarrow 3$.
As $1 \leftrightarrow 2 \leftrightarrow 3$, *this is an irreducible.*

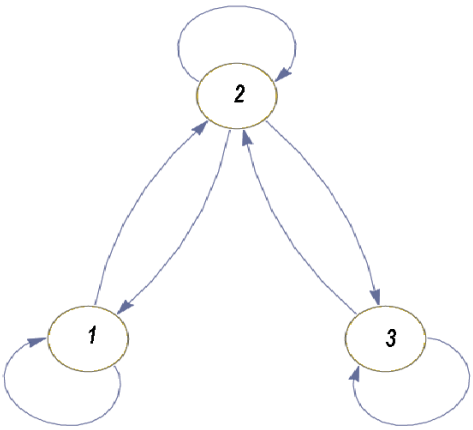

Figure 3.7: Transition graph for a Markov chain in Example 3.3.1

Example 3.3.2. *Consider a Markov chain with a state space* $\Omega = \{1, 2, 3, 4\}$ *and a transition matrix*

$$\mathrm{T} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

which has a transition graph as follows:

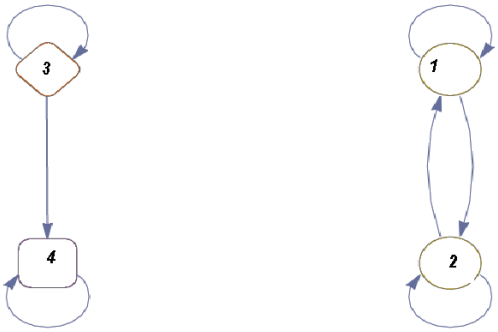

Figure 3.8: Transition graph for a Markov chain in Example 3.3.2

$p(1,\ 2) = \frac{1}{2} > 0$ *and* $p(2, 1) = \frac{1}{2} > 0$ *then* $1 \leftrightarrow 2$.

$p(3,\ 4) = \frac{3}{4} > 0$ *and* $p(4,3) = 0$ *then* $3 \nleftrightarrow 4$.
$\nexists n \geq 1$ *such that* $p^n(3,4) > 0$, *then this chain is not irreducible.*

Definition 3.3.4. [19] A state i is an ***absorbing*** state if $p(i,i) = 1$.

Definition 3.3.5. [8] A Markov chain with a state space $\Omega = \{x_0, x_1, ...\}$ and a transition matrix T is ***aperiodic*** if

$$\gcd\{n \geq 1 : p^n(i,\ j) > 0\} = 1, \qquad \forall i, j \in \Omega. \tag{3.10}$$

Otherwise the Markov chain is said to be ***periodic***.

Lemma 3.3.2. [23] *If* $p(i,i) > 0$, *then* i *has period* 1.

Proof. If $p(i,i) > 0$, then $1 \in \{n \geq 1 : p^n(i,j) > 0\}$,
so $\gcd\{n \geq 1 : p^n(i,j) > 0\} = 1$, then i has period 1. □

Definition 3.3.6. *[3]* A Markov chain is ***ergodic*** if it is both irreducible and aperiodic.

Example 3.3.3. *The Markov chain in Example* (3.2.1) *is aperiodic. Since*

$$\begin{aligned} d(1) &= \gcd\{n \geq 1 : p^n(1,\ 1) > 0\} \\ &= \gcd\{1,\ 2,\ 3, ...\} \\ &= 1, \end{aligned}$$

$$\begin{aligned} d(2) &= \gcd\{n \geq 1 : p^n(2,\ 2) > 0\} \\ &= \gcd\{1,\ 2,\ 3, ...\} \\ &= 1, \end{aligned}$$

and $1 \leftrightarrow 2$, *then the Markov chain is irreducible.*
Therefore the Markov chain is ergodic.

Let $A(i)$: denote the set of all accessible states from i

$$A(i) \;=\; \{j \in \Omega : i \to j\}. \tag{3.11}$$

Theorem 3.3.3. [17] *If $j \in A(i)$ and $k \in A(j)$, then $k \in A(i)$.*

Proof. Since $j \in A(i)$, then $i \to j$.

Since $k \in A(j)$, then $j \to k$, by Theorem (3.3.1), $i \to k$.

Therefore $k \in A(i)$. □

Definition 3.3.7. *[17]* For a Markov chain $\{X_0, X_1, ...\}$ with a state space $\Omega = \{x_0, x_1, ...\}$. A state $i \in \Omega$ is ***recurrent*** if

$$j \;\to\; i, \qquad \forall\, j \in A(i). \tag{3.12}$$

If a state is not recurrent, we say it is ***transient***.

Example 3.3.4. [17] *Consider a Markov chain with a state space $\Omega = \{1, 2, 3, 4, 5\}$ and a transition matrix*

$$\mathrm{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

Since $5 \to 5$, then $A(5) = \{5\}$,

since $2 \to 1$, $2 \to 2$, $2 \to 3$, $2 \to 4$ and $2 \to 5$, then $A(2) = \Omega$,

since $3 \to 1$, $3 \to 2$, $3 \to 3$, $3 \to 4$ and $3 \to 5$, then $A(3) = \Omega$,

since $4 \to 1$, $4 \to 2$, $4 \to 3$, $4 \to 4$ and $4 \to 5$, then $A(4) = \Omega$,

since $1 \to 1$, then $A(1) = \{1\}$.

Now

since $5 \in A(5)$ and $5 \to 5$, then 5 is recurrent,

since $1 \in A(4)$ and $1 \nrightarrow 4$, then 4 is transient,

since $1 \in A(3)$ and $1 \nrightarrow 3$, then 3 is transient,

since $1 \in A(2)$ and $1 \nrightarrow 2$, then 2 is transient,

since $1 \in A(1)$ and $1 \to 1$, then 1 is recurrent.

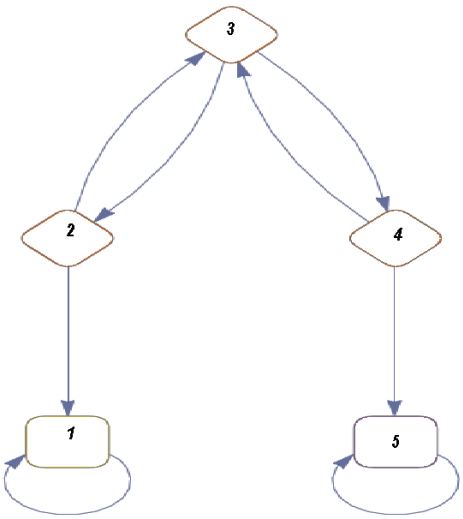

Figure 3.9: Transition graph for a Markov chain in Example 3.3.4

Theorem 3.3.4. [17] *If i is recurrent and $j \in A(i)$, then $A(j) = A(i)$.*

Proof. Let i is recurrent and $j \in A(i)$. We want show $A(j) \subset A(i)$ and $A(i) \subset A(j)$.

First let $k \in A(j)$, then $j \to k$ and since $j \in A(i)$, then $i \to j$.

Then by Theorem (3.3.1), $i \to k$ and $k \in A(i)$. Therefore $A(j) \subset A(i)$.

Second let $k \in A(i)$, then $i \to k$.

Since i is recurrent and $j \in A(i)$ then by Definition (3.3.7), j is recurrent, therefore $j \to i$.

By Theorem (3.3.1), if $j \to i$ and $i \to k$ then $j \to k$ and $k \in A(j)$.

Therefore $A(i) \subset A(j)$. Hence $A(j) = A(i)$. □

Theorem 3.3.5. [19] *If C is a finite closed and irreducible set, then all states in C are recurrent.*

Proof. Since C is a finite closed set, then if $x \in C$ and $z \notin C$ then $p(x, z) = 0$.

Since C is irreducible set, then $\forall x, y \in C$, $x \leftrightarrow y$.

Therefore $\forall x \in C$, x is recurrent. □

Let us introduce some definitions that will help us identify recurrent and transient states.

Define T_i = Time for first visit to i given $X_0 = i$:

$$T_i = \min\{n \geq 1 : X_n = i\}. \tag{3.13}$$

T_i known as the hitting time with $T = \infty$ if no such times exists.

Define f_{ii}^n be probability of first recurrence to i at the n^{th} step:

$$f_{ii}^n = P\{X_n = i, X_{n-1} \neq i, ..., X_1 \neq i | X_0 = i\}. \tag{3.14}$$

Note:

$$\begin{aligned} f_{ii}^n &= P\{X_n = i, X_{n-1} \neq i, ..., X_1 \neq i | X_0 = i\} \\ &= P\{T_i = n | X_0 = i\}. \end{aligned}$$

Define f_i be probability of recurrence to i:

$$f_i = \sum_{n=1}^{\infty} f_{ii}^n. \tag{3.15}$$

In the words,

$$\begin{aligned} f_i &= probability\ of\ ever\ returning\ to\ i \\ &= P\{T_i < \infty | X_0 = i\}. \end{aligned}$$

If $f_i < 1$ then

$$\begin{aligned} 1 - f_i &= probability\ of\ never\ returning\ to\ i \\ &= P\{T_i = \infty | X_0 = i\}. \end{aligned}$$

Definition 3.3.8. [19] A state i is a ***recurrent*** state if $f_i = 1$.

Definition 3.3.9. [19] A state i is a ***transient*** state if $f_i < 1$.

Let us now compute the expected number of visits to i (i.e., the times, including time 0, when the chain is at i).

$$P(exactly\ n\ visits\ to\ i|X_0 = i) \quad = \quad f_i^{n-1}(1 - f_i). \tag{3.16}$$

This formula says that the number of visits to i is a Geomtric $(1 - f_i)$ random variable so its expectation equals

$$E(number\ of\ visits\ to\ i|X_0 = i) \quad = \quad \frac{1}{1 - f_i}. \tag{3.17}$$

Theorem 3.3.6. [19] *If N is number of visits to i given $X_0 = i$ then*

$$E[N|X_0 = i] \quad = \quad \sum_{n=0}^{\infty} p^n(i, i). \tag{3.18}$$

Proof. Let $N = \sum_{n=0}^{\infty} I_n$, where

$$I_n = \begin{cases} 1, & X_n = i; \\ 0, & \text{otherwise.} \end{cases}$$

Since

$$\begin{aligned} P\{I_n = 1|X_0 = i\} &= P\{X_n = i|X_0 = i\} \\ &= p^n(i, i), \end{aligned}$$

then, we have

$$\begin{aligned} E[N|X_0 = i] &= E[\sum_{n=0}^{\infty} I_n|X_0 = i] \\ &= \sum_{n=0}^{\infty} E[I_n|X_0 = i] \\ &= \sum_{n=0}^{\infty} P\{X_n = i|X_0 = i\} \\ &= \sum_{n=0}^{\infty} p^n(i, i). \end{aligned}$$

So

$$E[N|X_0 = i] = \sum_{n=0}^{\infty} p^n(i, i).$$

□

Theorem 3.3.7. [19] *i is recurrent if and only if*

$$\sum_{n=0}^{\infty} p^n(i,\, i) = \infty. \tag{3.19}$$

Proof. Since

$$\begin{aligned} E[N|X_0 = i] &= \sum_{n=0}^{\infty} p^n(i,i) \\ &= \frac{1}{1-f_i}, \end{aligned}$$

then

$$\begin{aligned} \sum_{n=0}^{\infty} p^n(i,i) &= \infty \\ &iff \quad f_i = 1, \end{aligned}$$

which is the defintion of recurrence. □

Theorem 3.3.8. [19] *If state i is recurrent and state j communicate with state i, then state j is also recurrent.*

Proof. Let $\{X_0,\, X_1, ...\}$ be a Markov chain with a state space $\Omega = \{x_0,\, x_1, ...\}$ and $i, j \in \Omega$,

since i and j communicate then $\exists n, m > 0$ such that $p^n(i,j) > 0$ and $p^m(j,i) > 0$.

Since i is recurrent, then $\sum_{k=1}^{\infty} p^k(i,\, i) = \infty$.

For $k \geq 1$, by applying Chapman-Kolmogorov equations repeatedly yields:

$$\begin{aligned} p^{n+m+k}(j,\, j) &= \sum_{l\in\Omega} p^m(j,\, l)\, p^{n+k}(l,\, j) \\ &\geq p^m(j,\, i)\, p^{n+k}(i,\, j) \\ &= p^m(j,\, i) \sum_{l\in\Omega} p^k(i,\, l)\, p^n(l,\, j) \\ &\geq p^m(j,\, i)\, p^k(i,\, i)\, p^n(i,\, j). \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{k=1}^{\infty} p^{n+m+k}(j,\, j) &\geq \sum_{k=1}^{\infty} p^m(j,\, i)\, p^k(i,\, i)\, p^n(i,\, j) \\ &= p^m(j,\, i)\, p^n(i,\, j) \sum_{k=1}^{\infty} p^k(i,\, i). \end{aligned}$$

Since $p^n(i,j) > 0$, $p^m(j,i) > 0$ and $\sum_{k=1}^{\infty} p^k(i,i) = \infty$, then

$$\sum_{k=1}^{\infty} p^{n+m+k}(j,\, j) \;\; = \;\; \infty.$$

Hence j is also recurrent state. □

3.4 Stationary Distributions

In this section, we introduce the definition of stationary distribution for Markov chain and give some examples. Then we present the definition of doubly stochastic and give example.

Definition 3.4.1. *[3]* Let $\{X_n : \mathrm{n} \geq 0\}$ be a Markov chain with a state space Ω and a transition matrix T. $\pi = (\pi_i,\, i \in \Omega\ )$ is said to be a ***Stationary Distribution*** for Markov chain if it satisfies:

1. $\pi_i \geq 0, \qquad \forall\ i \in \Omega.$
2. $\sum_{i\in\Omega} \pi_i = 1.$
3. $\pi\, \mathrm{T} = \pi$, i.e., $\pi_j = \sum_{i\in\Omega} \pi_i\, p(i,j), \qquad \forall\ j \in \Omega.$

Example 3.4.1. *Consider a Markov chain with a state space* $\Omega = \{0,1,2,3\}$ *and a transition matrix*

$$\mathrm{T} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

Let $\pi = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$. *The equation* $\pi\ \mathrm{T} = \pi$ *says*

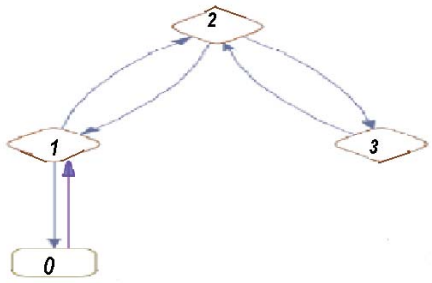

Figure 3.10: Transition graph for a Markov chain in Example 3.4.1

$$\begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$$

which translates into four equations

$$\pi_0 = \tfrac{1}{3}\pi_1,$$

$$\pi_0 + \tfrac{2}{3}\pi_2 = \pi_1 \textit{ implies } \pi_2 = \pi_1,$$

$$\tfrac{2}{3}\pi_1 + \pi_3 = \pi_2,$$

$$\tfrac{1}{3}\pi_2 = \pi_3 \textit{ implies } \pi_3 = \tfrac{1}{3}\pi_1.$$

If we add up the four equations we get

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 \;=\; \pi_0 + \pi_1 + \pi_2 + \pi_3.$$

Since

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 \;=\; 1,$$

then

$$\frac{1}{3}\pi_1 + \pi_1 + \pi_1 + \frac{1}{3}\pi_1 \;=\; 1,$$

and

$$\pi_1 = \tfrac{3}{8},$$

therefore

$$\begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}.$$

To check this we note that

$$\begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

Definition 3.4.2. [19] A transition matrix T is said to be ***doubly stochastic*** if its columns sum to 1, or in symbols

$$\sum_i p(i,\, j) \;=\; 1. \tag{3.20}$$

Theorem 3.4.1. [19] *If* T *is a doubly stochastic transition matrix for a Markov chain with a state space* $\Omega = \{1, 2, 3, ..., N\}$*, then the uniform distribution,* $\forall x \in \Omega$*,* $f(x) = \frac{1}{N}$*, is a stationary distribution.*

Proof. Let T is a doubly stochastic transition matrix for a Markov chain with N states, $y \in \Omega$ and $f(x) = \frac{1}{N}$, $\forall x \in \Omega$.

$$\begin{aligned} \sum_{x\in\Omega} f(x)\, p(x,\, y) &= \frac{1}{N} \sum_{x\in\Omega} p(x,\, y) \\ &= \frac{1}{N} \\ &= f(y). \end{aligned}$$

Therefore $\forall x \in \Omega$, $f(x) = \frac{1}{N}$ is a stationary distribution. □

Example 3.4.2 (Symmetric reflecting random walk on the line)**.** [23] *Let* $\Omega = \{1, 2, ..., L\}$ *is a state space, the chain goes to the right or left at each step with probability* $\frac{1}{2}$*, subject to the rules that if it tries to go to the left from* 1 *or to the right from* L *it stays put. Let* $L = 5$*, then the transition matrix is*

$$\mathrm{T} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix},$$

It is clear in the example $N = 5$*, that each column adds up to* 1.

Therefore T *is doubly stochastic.*

So the stationary distribution is uniform, $\pi_i = \frac{1}{5}$, $\forall i \in \Omega$.

Therefore

$$\pi = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}.$$

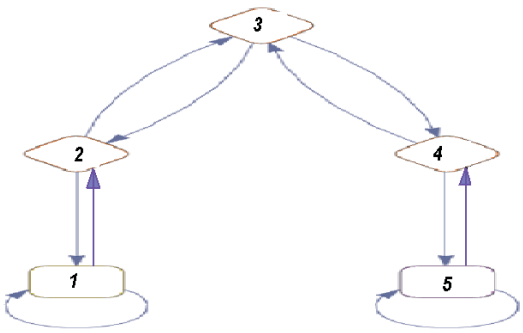

Figure 3.11: Transition graph for a Markov chain in Example 3.4.2

3.5 Detailed Balance and Time Reversal

Definition 3.5.1. [8] A stationary Markov chain with a transition matrix T and stationary distribution π is called ***reversible***

if for arbitrary $t \geq 0$ and $x_0, x_1, ..., x_t \in \Omega$,

$P(X_0 = x_0,\ X_1 = x_1,\ ...,\ X_{t-1} = x_{t-1},\ X_t = x_t) = P(X_t = x_0,\ X_{t-1} = x_1,\ ...,\ X_1 = x_{t-1},\ X_0 = x_t)$.

Definition 3.5.2. [8] A probability distribution π on the state space Ω is reversible for the Markov chain $\{X_0, X_1, ...\}$ with transition matrix T if

$$\pi_i p(i,j) \;=\; \pi_j p(j,i), \qquad \forall i,\, j \in \Omega. \tag{3.21}$$

Which is also known as ***detailed balance equation***.

Example 3.5.1 (Brand preference)**.** [19] *Suppose there are three types of laundry detergent,* $1, 2,$ *and* 3*, and let* X_n *be the brand chosen on the nth purchase. Customers who try these brands are satisfied and choose the same thing again with probabilities* $0.8, 0.6,$ *and* 0.4 *respectively.*
When they change they pick one of the other two brands at random.

The transition matrix is

$$\mathrm{T} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}.$$

Let $\pi = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$.
The equation $\pi\ \mathrm{T} = \pi$ *says*

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

which translates into three equations

$$\begin{aligned} 0.8\pi_1 + 0.2\pi_2 + 0.3\pi_3 &= \pi_1, \\ 0.1\pi_1 + 0.6\pi_2 + 0.3\pi_3 &= \pi_2, \\ 0.1\pi_1 + 0.2\pi_2 + 0.4\pi_3 &= \pi_3. \end{aligned}$$

Note that if we add up the three equations we get

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

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

$$\begin{aligned} -0.2\pi_1 + 0.2\pi_2 + 0.3\pi_3 &= 0, \\ 0.1\pi_1 - 0.4\pi_2 + 0.3\pi_3 &= 0, \\ \pi_1 + \pi_2 + \pi_3 &= 1. \end{aligned}$$

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

$$\begin{aligned} 0.5\pi_1 + 0.1\pi_2 &= 0.3, \\ 0.2\pi_1 + 0.7\pi_2 &= 0.3. \end{aligned}$$

Multiplying the first equation by 0.2 *and the second equation by* -0.5 *gives*

$$\begin{aligned} 0.1\pi_1 + 0.02\pi_2 &= 0.06, \\ -0.1\pi_1 - 0.35\pi_2 &= -0.15. \end{aligned}$$

Adding the first equation and second equation gives $\pi_2 = \frac{3}{11}$, *substituting this in the first equation gives*

$$\pi_1 = \tfrac{6}{11},$$

since $\pi_1 + \pi_2 + \pi_3 = 1$, *then*

$$\pi_3 = \tfrac{2}{11}.$$

Therefore the stationary distribution is given by:

$$\pi = \begin{pmatrix} \frac{6}{11} & \frac{3}{11} & \frac{2}{11} \end{pmatrix}.$$

Since

$$\begin{aligned}\pi_1 p(1,2) &= \pi_2 p(2,1) = \frac{6}{110},\\ \pi_1 p(1,3) &= \pi_3 p(3,1) = \frac{6}{110},\\ \pi_2 p(2,3) &= \pi_3 p(3,2) = \frac{6}{110},\end{aligned}$$

then the detailed balance equation (reversibility) holds.

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

Example 3.5.2 (Social mobility)**.** [19] *Let X_n be a family's social class in the n^{th} generation, which we assume is either $1 =$ lower, $2 =$ middle, or $3 =$ upper. In our simple version of sociology, changes of status are a Markov chain with the following transition matrix*

$$\mathrm{T} = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}.$$

Let $\pi = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$. The equation $\pi\ \mathrm{T} = \pi$ 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

$$\begin{aligned}0.7\pi_1 + 0.3\pi_2 + 0.2\pi_3 &= \pi_1,\\ 0.2\pi_1 + 0.5\pi_2 + 0.4\pi_3 &= \pi_2,\\ 0.1\pi_1 + 0.2\pi_2 + 0.4\pi_3 &= \pi_3.\end{aligned}$$

Therefore the stationary distribution is given by

$$\pi = \left(\begin{array}{ccc} \frac{22}{47} & \frac{16}{47} & \frac{9}{47} \end{array} \right).$$

To chick this we note that

$$\left(\begin{array}{ccc} \frac{22}{47} & \frac{16}{47} & \frac{9}{47} \end{array} \right) \left(\begin{array}{ccc} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{array} \right) = \left(\begin{array}{ccc} \frac{22}{47} & \frac{16}{47} & \frac{9}{47} \end{array} \right).$$

Since

$$\begin{aligned} \pi_1 p(1,2) &= \frac{22}{47}\frac{2}{10} \\ &= \frac{44}{470}, \end{aligned}$$

and

$$\begin{aligned} \pi_2 p(2,1) &= \frac{16}{47}\frac{3}{10} \\ &= \frac{48}{470}, \end{aligned}$$

then

$$\pi_1 \mathrm{p}(1,2) \neq \pi_2 \mathrm{p}(2,1).$$

Therefore the reversibility fails.

Theorem 3.5.1 (Detailed Balance Test)**.** [13] *If the probability distribution π is reversible for a Markov chain, then it is also a stationary distribution for the chain.*

Proof. Since π is reversible, we have

$$\pi_i \, p(i,\, j) = \pi_j \, p(j,\, i), \qquad \forall i, j \in \Omega.$$

For fixed $i \in \Omega$, we sum this equation with respect to $j \in \Omega$ to get

$$\sum_j \pi_i \, p(i,\, j) = \sum_j \pi_j \, p(j,\, i).$$

But the left hand side equals $\pi_i \sum_j p(i,j) = \pi_i$, since rows sum to 1.
Thus

$$\pi_i \;=\; \sum_j \pi_j \, p(j,\, i), \qquad \forall i, j \in \Omega,$$

and this implies $\pi \, \mathrm{T} = \pi$, which makes π a stationary distribution. □

Example 3.5.3. *Consider a Markov chain with a state space $\Omega = \{1, 2\}$ and a transition matrix*

$$\mathrm{T} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix},$$

Let

$$\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Since

$$\pi_1 \, p(1,\, 2) \;=\; \pi_2 \, p(2,\, 1) = \frac{1}{4},$$

then the reversibility holds, by previous theorem π is stationary distribution.
To chick this we must compute the stationary distribution.
Let $\pi = \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix}$. We want solve

$$\begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix}.$$

Multiplying gives two equations

$$\begin{aligned} 0.5\pi_1 + 0.5\pi_2 &= \pi_1, \\ 0.5\pi_1 + 0.5\pi_2 &= \pi_2. \end{aligned}$$

Both equations reduce to $\pi_1 = \pi_2$.
Since $\pi_1 + \pi_2 = 1$, we must have $2\pi_2 = 1$ which implies $\pi_1 = \pi_2 = \frac{1}{2}$. Hence

$$\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Chapter 4

Markov Chain Monte Carlo Methods

4.1 Introduction

In this chapter, we will study some roles of Markov Chain Monte Carlo methods in Bayesian inference. Suppose that X has a distribution with a likelihood function $f(x|\theta)$ and a prior density for θ given by $f(\theta)$. As we saw in Section (1.2), Bayes's Theorem relates the posterior $f(\theta|x)$ to the prior via the formula:

$$f(\theta|x) \quad \propto \quad f(x|\theta)f(\theta), \tag{4.1}$$

where the constant of proportionality (normalizing constant) is given by:

$$z \quad = \quad \int f(x|\theta)f(\theta)d\theta. \tag{4.2}$$

The Bayesian inference proceeds from posterior distribution

$$\pi(x) = f(\theta|x) \quad = \quad \frac{f(x|\theta)f(\theta)}{z}. \tag{4.3}$$

The denominator z is often very difficult to compute. We can use Markov Chain Monte Carlo (MCMC) methods to sample from a complicated or high dimensional posterior distribution. See [16].

4.2 Markov Chain Monte Carlo Algorithms

The general idea of MCMC methods in Bayesian inference is a class algorithms for sampling from posterior distribution $\pi(x)$ based on constructing a Markov chain $\{X_0, X_1, X_2, ...\}$ having a state space Ω and has posterior distribution as its stationary distribution.

Markov chains are usually constructed from a transition kernal $p(X_t, X_{t+1})$. A Markov chain generates a sequence of x values, denoted by $(x_0, x_1, x_2, ...)$ in such a way that as $n \to \infty$, we can guarantee that $x_n \sim \pi(x)$.

We present in the following subsections common MCMC methods; namely, the Metropolis-Hastings algorithms and the Gibbs sampler. See [24].

4.2.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is the most popular example of the MCMC methods. Suppose we have posterior distribution $\pi(x)$, if we want to sample from $\pi(x)$, then we construct a Markov chain whose stationary distribution is $\pi(x)$, and run a Markov chain long enough and then use Metropolis-Hastings algorithm. See [18].

The algorithm proceeds as follows:

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

$$\alpha = \min\left\{1, \frac{\pi(x^*)q(x^*, x)}{\pi(x)q(x, x^*)}\right\}. \tag{4.4}$$

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

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

(we accept the candidate point x^* with probability α; i.e., we accept x^* if $u \leq \alpha$).

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

If the proposal distribution is symmetric,

$$q(x, y) = q(y, x), \tag{4.5}$$

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

$$\alpha = \min\left\{1, \frac{\pi(x^*)}{\pi(x)}\right\}. \tag{4.6}$$

Theorem 4.2.1. [12] *The Metropolis algorithm produce a Markov chain* $\{X_0, X_1, X_2, ...\}$ *which is reversible with respect to stationary distribution* $\pi(x)$.

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

Set $P(x, y) = q(x, y)\alpha$ to construct the transition probabilities.

We must show that $\pi(x)P(x, y) = \pi(y)P(y, x)$.

Obviously this holds if $x = y$.

We will consider $x \neq y$, then

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

and

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

Hence

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

Therefore, the chain is reversible with respect to stationary distribution $\pi(x)$. $\square$

This example explains the Metropolis algorithm manually:

Example 4.2.1. [9] *Write the Metropolis algorithm for obtaining samples from the posterior distribution* $\pi(x) = N(5, 1.5^2)$.
The posterior distribution is given by:

$$\begin{aligned}
\pi(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \frac{1}{1.5\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2(1.5^2)}} \\
&\propto e^{-\frac{2(x-5)^2}{9}}.
\end{aligned}$$

We can select the proposal distribution $q(x, x^*) = N(x, 1)$, *which is symmetric distribution.*
The acceptance probability α *is given by:*

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi(x^*)}{\pi(x)}\right\} \\
&= \min\left\{1, \frac{e^{-\frac{2(x^*-5)^2}{9}}}{e^{-\frac{2(x-5)^2}{9}}}\right\}.
\end{aligned}$$

The Metropolis algorithm proceeds as follows:

- *Select starting point* $x_0 = 0$, *generate* $x^* = 0.2893$ *from* $N(x_0 = 0, 1)$, *and* $u = 0.571$ *from uniform*$(0, 1)$.
  Then the acceptance probability

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{e^{-\frac{2(0.2893-5)^2}{9}}}{e^{-\frac{2(0-5)^2}{9}}}\right\} \\
&= \min\{1, 1.3198\} \\
&= 1.
\end{aligned}$$

Since $u = 0.571 < \alpha = 1$, *then the Metropolis algorithm accepts* $x^* = 0.2893$.
Therefore $x_1 = 0.2893$ *with probability* $\alpha = 1$.

- *Generate* $x^* = 0.4462$ *from* $N(x_1 = 0.2893, 1)$,
  and $u = 0.9801$ *from uniform*$(0, 1)$.
  Then the acceptance probability

$$\begin{aligned} \alpha &= \min\left\{1, \frac{e^{-\frac{2(0.4462-5)^2}{9}}}{e^{-\frac{2(0.2893-5)^2}{9}}}\right\} \\ &= \min\{1, 1.1544\} \\ &= 1. \end{aligned}$$

  Since $u = 0.9801 \leq \alpha = 1$, *then the Metropolis algorithm accepts the point* $x^* = 0.4462$.
  Therefore $x_2 = 0.4462$ *with probability* $\alpha = 1$.

- *Generate* $x^* = -1.8602$ *from* $N(x_2 = 0.4462, 1)$,
  and $u = 0.211$ *from uniform*$(0, 1)$.
  Then the acceptance probability

$$\begin{aligned} \alpha &= \min\left\{1, \frac{e^{-\frac{2(-1.8602-5)^2}{9}}}{e^{-\frac{2(0.4462-5)^2}{9}}}\right\} \\ &= \min\{1, 0.0743\} \\ &= 0.0743. \end{aligned}$$

  Since $u = 0.211 > \alpha = 0.0743$, *then the Metropolis algorithm rejects the point* $x^* = -1.8602$.

- *Generate* $x^* = 0.6989$ *from* $N(x_2 = 0.4462, 1)$,
  and $u = 0.003$ *from uniform*$(0, 1)$.

Then the acceptance probability

$$\begin{aligned}\alpha &= \min\left\{1, \frac{e^{-\frac{2(0.6989-5)^2}{9}}}{e^{-\frac{2(0.4462-5)^2}{9}}}\right\} \\ &= \min\{1, 1.2473\} \\ &= 1.\end{aligned}$$

Since $u = 0.003 \leq \alpha = 1$*, then the Metropolis algorithm accepts the point* $x^* = 0.6989$*.*

Therefore $x_3 = 0.6989$ *with probability* $\alpha = 1$*.*

- *Generate* $x^* = -0.8729$ *from* $N(x_3 = 0.6989, 1)$*,*
  and $u = 0.157$ *from uniform*$(0, 1)$*.*
  Then the acceptance probability

$$\begin{aligned}\alpha &= \min\left\{1, \frac{e^{-\frac{2(-0.8729-5)^2}{9}}}{e^{-\frac{2(0.6989-5)^2}{9}}}\right\} \\ &= \min\{1, 0.2061\} \\ &= 0.2061.\end{aligned}$$

  Since $u = 0.157 \leq \alpha = 0.2061$*, then the Metropolis algorithm accepts the point* $x^* = -0.8729$*.*

  Therefore $x_4 = -0.8729$ *with probability* $\alpha = 0.2061$*.*

Therefore $x_1 = 0.2893, x_2 = 0.4462, x_3 = 0.6989, x_4 = -0.8729, \ldots\ldots$ *And continue as above for (say)* 200 *values for* x

4.2.2 Gibbs Sampler

The Gibbs sampler is one way of MCMC methods, which help us to generate samples from joint (posterior) distributions. In this method, the samples do not generate directly from the joint (posterior) distribution, but generate from the conditional distributions derived from the joint (posterior) distribution. See [1].

To introduce the Gibbs sampler, let π be a joint (posterior) distribution of a bivariate random vector (X, Y). Let $\pi(X|Y)$ be the conditional probability distribution of X given Y. Similarly, let $\pi(Y|X)$ be the conditional probability distribution of Y given X. See [4].

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

Start with some $X_0 = x_0$,

$$X_k \sim \pi(X|Y_{k-1}), \qquad for\ k = 1, 2, ... \tag{4.7}$$

$$Y_k \sim \pi(Y|X_k), \qquad for\ k = 0, 1, 2, ... \tag{4.8}$$

The next example explains the Gibbs sampler manually:

Example 4.2.2. [13] *Suppose the joint (posterior) distribution of* $x = 0, 1, ..., n$ *and* $0 \leq y \leq 1$ *is given by:*

$$\pi(x, y) = \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}. \tag{4.9}$$

Now we need to calculate the marginal distribution of x *and the marginal distribution of* y *as follows:*

The marginal distribution of x *is given by:*

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

The marginal distribution of y is given by:

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

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

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

Thus,

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

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

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

Thus,

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

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

$$x_k \sim Bi(n, y_{k-1}), \qquad for\ k = 1, 2, ... \tag{4.10}$$

$$y_k \sim Be(x_k + \alpha, n - x_k + \beta), \qquad for\ k = 0, 1, 2, ... \tag{4.11}$$

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

- *Start with* $x_0 = 2$ *and use it to obtain* y_0 *from* (4.11)*:*

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

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

- x_1 *is obtained from* (4.10)*:*

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

  which gives $x_1 = 2$.
  y_1 *is obtained from* (4.11)*:*

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

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

- x_2 *is obtained from* (4.10)*:*

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

giving $x_2 = 3$.

And y_2 *is obtained from* (4.11)*:*

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

giving $y_2 = 0.4735$.

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

- x_3 *is obtained from* (4.10)*:*

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

giving $x_3 = 6$.

And y_3 *is obtained from* (4.11)*:*

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

giving $y_3 = 0.6338$.

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

- x_4 *is obtained from* (4.10)*:*

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

giving $x_4 = 4$.

And y_4 *is obtained from* (4.11)*:*

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

giving $y_4 = 0.4196$.

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

The Gibbs sequence within five terms:

$(2, 0.2379), (2, 0.1334), (3, 0.4735), (6, 0.6338), (4, 0.4196), ...$

Notation. Let us define $x_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, then $\pi(x_i|x_{-i}) = \pi(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm where the acceptance probability $\alpha = 1$.

To show $\alpha = 1$ let the proposal distribution be

$$\begin{aligned} q(x, y) &= \pi(y_i|x_{-i}), \ if \ y_{-i} = x_{-i}; \\ &= 0, \ otherwise. \end{aligned}$$

The acceptance probability is:

$$\begin{aligned} \alpha &= \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \\ &= \min\left\{1, \frac{\pi(y)\pi(x_i|y_{-i})}{\pi(x)\pi(y_i|x_{-i})}\right\} \\ &= \min\left\{1, \frac{\pi(y)\pi(x_i|x_{-i})}{\pi(x)\pi(y_i|y_{-i})}\right\} \\ &= \min\left\{1, \frac{\pi(y)\pi(x_i, x_{-i})\pi(y_{-i})}{\pi(x)\pi(y_i, y_{-i})\pi(x_{-i})}\right\} \\ &= \min\left\{1, \frac{\pi(y_{-i})}{\pi(x_{-i})}\right\} \\ &= \min\left\{1, \frac{\pi(x_{-i})}{\pi(x_{-i})}\right\} = 1. \end{aligned}$$

$\alpha = 1$ means the candidate point is always accepted.

The multivariate extension of the bivariate case is very straightforward. Suppose π is a probability distribution of a k-dimensional random vector $(X_1, X_2, \ldots, X_n)$. Let $\pi(X_i|X_{-i})$ denote the univariate conditional probability distribution of X_i given that $X_{-i} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Now starting with some initial value

for $X_0 = (x_{01}, x_{02}, \ldots, x_{0k})$, generate $X_1 = (x_{11}, x_{12}, \ldots, x_{1k})$ as follows:

$$\begin{aligned}
x_{11} \quad & from \quad \pi(x_1|x_{02}, x_{03}, \ldots, x_{0k}); \\
x_{12} \quad & from \quad \pi(x_2|x_{11}, x_{03}, \ldots, x_{0k}); \\
x_{13} \quad & from \quad \pi(x_3|x_{11}, x_{12}, x_{04}, x_{05}, \ldots, x_{0k}); \\
& \vdots \\
x_{1k} \quad & from \quad \pi(x_k|x_{11}, x_{12}, \ldots, x_{1(k-1)}).
\end{aligned}$$

Iteration of this process generates a sequence $(X_1, X_2, \ldots, X_n)$.

4.3 Simulation

This section has considered as an important application to the previous section. First, we present simulator which has used in simulate n values of Markov chains. Second, we introduce simulators which have used in simulate from posterior distributions by MCMC methods and give some examples. There are several computer programs which used in the simulation such as R program and Matlab program. In this section, we will use R program. See [23].

4.3.1 Markov chain simulators

There are many simulators which used in simulate n values of Markov chains. We intrest in using simulator when Markov chains have three states. Suppose we want simulate n values of a Markov chain having a transition matrix T, starting at x_1. We will use the a Markov chain Simulator by R program. See [11].
Let $x1 = x_1$ and $P =$ T, then a Markov chain simulator proceeds as follows:

$> MC.sim = function(n, P, x1)\{$
$+sim = as.numeric(n)$
$+m = ncol(P)$

$+if(missing(x1))\{$

$+sim[1] = sample(1:m, 1)$

$+\}else\{sim[1] = x1\}$

$+for(i\ in\ 2:n)\{$

$+newstate = sample(1:m, 1, prob = P[sim[i-1],])$

$+sim[i] = newstate$

$+\}$

$+sim$

$+\}.$

The next example illustrates the previous simulator:

Example 4.3.1 (Symmetric reflecting random walk on the line)**.** [11]
Let $E = \{1, 2, 3\}$ *is a state space, the chain goes to the right or left at each step with probability* $\frac{1}{2}$*, subject to the rules that if it tries to go to the left from* 1 *or to the right from* 3 *it stays put. Therefore the transition matrix is*

$$\mathrm{T} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

Suppose we want simulate 100 *values of a Markov chain having transition matrix* T*, starting at* $x_1 = 1$

Let $P = \mathrm{T}$*,* $n = 100$*, and* $x1 = x_1 = 1$*.*

First we must input P *by this commands:*

$> P = matrix(c(.5, .5, 0, .5, 0, .5, 0, .5, .5), nrow = 3)$

$> P$

```
     [,1] [,2] [,3]
[1,]  0.5  0.5  0.0
[2,]  0.5  0.0  0.5
[3,]  0.0  0.5  0.5
```

Second applying a Markov chain simulator as follows:

$$> MC.sim(100, P, 1)$$

A Markov chain simulator generates 100 *values of above Markov chain having transition matrix* T*, starting at* $x_1 = 1$

[1] 1 3 1 2 3 1 3 1 3 1 2 3 1 2 3 1 2 3 1 2 3 1 3 1 2 3 1 2 2 2 2 2 3 1 3 1 2

[38] 2 2 2 2 3 1 2 2 2 3 1 2 3 1 3 1 3 1 3 1 2 3 1 2 2 2 2 2 2 3 1 2 2 3 1 3 1

[75] 2 2 2 3 1 3 1 3 1 2 3 1 2 3 1 2 2 3 1 3 1 3 1 3 1 2.

4.3.2 MCMC simulators

Metropolis-Hastings simulator

Suppose we want simulate from a gamma distribution $gamma(a, b)$, using a Metropolis-Hastings algorithm with normal proposal distribution $N(\frac{a}{b}, b \times b)$.

See [5].

A Metropolis-Hastings simulator proceeds as follows:

$> gamm = function(n, a, b)$

{

$+mu = a/b$

$+sig = sqrt(a/(b * b))$

$+vec = vector("numeric", n)$

$+x = a/b$

$+vec[1] = x$

$+for(i\ in\ 2\ :\ n)\{$

$+can = rnorm(1, mu, sig)$

$+aprob = min(1, (dgamma(can, a, b)/dgamma(x, a, b))/(dnorm(can, mu, sig)/dnorm(x, mu, sig)))$

$+u = runif(1)$

$+if(u < aprob)$

$+x = can$

$+vec[i] = x$

+}

$+vec$

+}.

The next example illustrates the previous simulator:

Example 4.3.2. [11] *Simulate from a gamma distribution* $gamma(2.1, 2.7)$*, using a Metropolis-Hastings algorithm with normal proposal distribution* $N(\frac{2.1}{2.7}, 2.7 \times 2.7)$.

Suppose $a = 2.1$*,* $b = 2.7$ *and* $n = 100$.

Applying a Metropolis-Hastings simulator as follows:

$$> gamm(100, 2.1, 2.7)$$

A Metropolis-Hastings simulator generates 100 *values:*

[1] 0.77777778 0.04033136 1.59865052 0.10276840 1.65615693 0.39844543

[7] 1.19329359 1.39032911 0.61762474 0.61762474 0.15416545 0.15416545

[13] 0.15416545 0.15416545 0.98577761 0.18699319 1.22852452 0.52530612

[19] 0.21045406 0.21045406 0.21045406 0.91926053 0.81911859 0.92791276

[25] 0.92791276 0.29408300 0.29408300 0.39745840 1.23030988 2.08724101

[31] 0.85986875 0.49735538 0.26796353 0.57406627 0.86156001 1.31218679

[37] 1.42025610 0.30246252 0.25822552 1.54485201 1.54485201 1.50947673

[43] 0.90738689 0.52076450 0.76426186 0.92589754 1.43231521 0.78658503

[49] 0.43391773 0.43391773 0.60808470 1.15675410 1.38892389 0.77756480

[55] 1.88096690 1.19112237 1.06411757 1.02106514 1.33348184 1.54055319

[61] 0.63321668 0.63321668 0.63321668 1.09695253 1.94503370 1.94503370

[67] 1.94503370 0.41800188 0.41800188 1.06011683 0.24745596 0.24745596

[73] 0.24745596 0.56968803 0.56968803 0.56968803 0.64475849 1.05653320

[79] 0.76913772 0.50859869 0.38392521 0.48656030 0.59029012 0.26182881

[85] 0.26182881 0.26182881 0.91841417 1.20536851 1.16301334 1.21480367

[91] 1.24583157 0.68328299 0.06967883 0.68772152 1.33857770 0.79136414

[97] 0.79136414 0.73365685 0.73365685 0.92746035.

Metropolis simulator

Suppose we want simulate n values from a $N(0,1)$ with uniform proposal distribution $U(-alpha, alpha)$. See [11].

The Metropolis simulator proceeds as follows:

$> function(n, alpha)$

$+\{$

$+vec = vector("numeric", n)$

$+x = 0$

$+vec[1] = x$

$+for(i\ in\ 2\ :\ n)\{$

$+innov = runif(1, -alpha, alpha)$

$+can = x + innov$

$+aprob = min(1, dnorm(can)/dnorm(x))$

$+u = runif(1)$

$+if(u < aprob)$

$+x = can$

$+vec[i] = x$

$+\}$

$+vec$

$+\}.$

The next example illustrates the previous simulator:

Example 4.3.3. [5] *Simulate from a $N(0,1)$, using a Metropolis algorithm with uniform proposal distribution $U(-alpha, alpha)$.*

Suppose we want simulate $n = 20$ values from $N(0,1)$ with uniform proposal distribution $U(-1,1)$.

Applying a Metropolis simulator as follows:

$$> norm(20, 1)$$

A Metropolis simulator generates 20 *values:*

[1] 0.00000000 0.60956015 0.86738903 1.44297264 1.44297264 1.51615672

[7] 0.82477932 0.82477932 0.31640838 0.77994809 0.72612074 0.89362667

[13] 1.00641295 0.21015467 −0.25406782 0.24126156 −0.02220538 0.23093248

[19] −0.41267932 −0.0231050.

Gibbs simulator

Suppose we want simulate n values from a bivariate normal with zero mean and unit variance for the marginals. See [11].

A Gibbs simulator proceeds as follows:

$> gibbs = function(n, rho)$

$+\{$

$+mat = matrix(ncol = 2, nrow = n)$

$+x = 0$

$+y = 0$

$+mat[1,] = c(x, y)$

$+for(i\ in\ 2\ :\ n)\{$

$+x = rnorm(1, rho * y, sqrt(1 - rho^2))$

$+y = rnorm(1, rho * x, sqrt(1 - rho^2))$

$+mat[i,] = c(x, y)$

$+\}$

$+mat$

$+\}.$

The next example illustrates the previous simulator:

Example 4.3.4. [2] *Simulate $n = 20$ values from a bivariate normal with zero mean and unit variance for the marginals.*

Applying a Gibbs simulator as follows:

$$> gibbs(20, 0.6)$$

A Gibbs simulator generates 20 *values:*

$$(0.00000000, 0.00000000), (0.60712281, -0.32079906),$$
$$(0.06167020, 0.34382792), (0.49107092, -0.22236705),$$
$$(0.46743566, 1.45364568), (-0.35024231, -0.35127267),$$
$$(0.63847272, 1.12556478), (-0.89778783, 0.05477754),$$
$$(-0.51036004, 0.30365097), (1.71092130, 0.62262023),$$
$$(1.57486411, 0.62886917), (0.93638339, -0.16391671),$$
$$(0.45070434, 0.05379624), (-0.73757406, 0.76219532),$$
$$(0.38310972, -0.26616544), (-0.08234679, 1.40470064),$$
$$(0.24426366, 0.68352088), (-1.00522943, -0.76867532),$$
$$(-0.05544645, -0.24158201), (0.85683711, -1.30943650).$$

4.4 Conclusion

- The Bayesian inference proceeds from posterior distribution.
- We use the posterior distribution to draw conclusions about certain properties of the complicated distribution.
- The jeffreys priors are a good choice if a prior distributions $f(\theta)$ does not contain any information about a parameter θ.
- Bayes factor uses in the choice between two models and two hypotheses.
- The importance of Markov chains comes from this fact:

  The next state depend only on the current state.

- The general idea of MCMC methods in Bayesian inference is a class algorithms for sampling from posterior distribution $\pi(x)$ based on constructing a Markov chain having a posterior distribution as its stationary distribution.
- The Gibbs sampler is a special case of the Metropolis-Hastings algorithm where the acceptance probability equals to 1.
- The Gibbs sampler is one way of MCMC methods, which help us to generate samples from joint (posterior) distributions. In this method, the samples do not generate directly from the joint (posterior) distribution, but generate from the conditional distributios derived from the joint (posterior) distribution.
- There are many simulators which used in simulate n values of finite Markov chains.
- There are many simulators which help us to generate samples from posterior distributions by MCMC methods.

Refrences

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

[2] Brian S. Everitt, An R and S-PLUS Companion to Multivariate, Analysis, springer (2006).

[3] Can Mert, Bayesian inference in Schizophrenia research, Unpublished master's thesis, Department of Numerical Analysis and Computer Science Royal Institute of Technology, Stockholm, Sweden (2002).

[4] Casella, G. and George, E. I. Explaining the Gibbs sampler, Am. Stat. 46(1992), 167-174.

[5] Christian P. Robert, George Casella, Introducing Monte Carlo Methods with R, Springer (2009).

[6] D. Dawson, Introduction to Markov Chains, Mcgill Unversity (1970).

[7] D. Revuz, Markov Chain, Elsevier Science Publishers (2005).

[8] David A. Levin, Yuval Peres and Elizabeth L. Wilmer, Markov Chains and Mixing Times, University of Oregon (2006).

[9] Dr S. K. Sahu, Lecture Notes: Bayesian Methods, Southampton Unversity (2001).

[10] Eliane Regina Rodrigues and Jorge Alberto Achcar, Applications of Discrete-Time Markov Chains and Poisson Process to Air Pollution Modeling and Studies, Springer (2013).

[11] Eric A. Suess and Bruce E. Trumbo, Introduction to Probability Simulation and Gibbs Sampling with R, Springer (2012).

[12] Evans, Michael J. and Rosenthal, J. S. probability and statistics, The science of Uncertainty, W. H. Freeman and Company, NewYork (2003).

[13] George Casella and Roger L. Berger, Statistical Inference, the Wadsworth Group(2002).

[14] Gilks, W.R., Richardson, S. and Spiegelhater, D.J. Markov Chain Monte Carlo In Practice, Chapman and Hall, London (1996).

[15] Jayanta K. Ghosh, An Introduction to Stochastic Processes, Springer (2006).

[16] O. Haggstrom, Finite Markov Chains and Algorithmic Applications, Cambridge University Press (2000).

[17] Pejman Mahboubi, Markov Chains, Recurrence, Transience, Periodicity, and Steady State (2005),
http://www.docstoc.com/docs/155812664/LiveMeeting11.

[18] Peter Lenk, Bayesian Inference and Markov Chain Monte Carlo, Michigan University (2001).

[19] Richard Durret, Essential of Stochastic Processes, Springer (2012).

[20] Richard J. Larsen and Morris L. Marx, An Introduction to Mathematical Statistic and Its Applications, Pearon Education, Inc (2006).

[21] Robert V. Hogg and Allen T. Craig, Introduction to Mathematical Statistics, Macmillan Publishing co., Inc (1978)

[22] Ronald Harshbarger, James J. Reynolds, Mathematical Applications for the Management, Life, and Social Sciences, Tenth Edition, Cengage Learning(2012).

[23] Saeed Ghahraman, Fundamentals of Probability with Stochastic Processes, Pareson Education, Inc (2005).

[24] Simon Jackman, Bayesian Analysis for the Social Sciences, John Wiley & Sons (2009).

[25] W.K. Hastings, Monte Carlo sampling using Markov chains and their Applications Biometrika, 57(1970), 97-109.

[26] W S Kendall, F Liang and J-S Wang, Markov Chain Monte Carlo Innovations and applications, World Scientic Publishing Co. Pte. Ltd., (2005).