MCMC Methods -An Application in Genetics

by

Dr. Geetha Antony Pullen

Final Report

of

MINOR RESEARCH PROJECT

MRP(S)-889/10-11/KLKA020/UGC-SWRO

DEPARTMENT OF STATISTICS
MARY MATHA ARTS AND SCIENCE COLLEGE
MANANTHAVADY, KERALA, INDIA.

1 Introduction

Over the last ten years, the introduction of computer intensive statistical methods has opened new horizons concerning the probability models that can be fitted to genetic data, the scale of the problems that can be tackled and the nature of the questions that can be posed. In particular, the application of Bayesian and likelihood methods to statistical genetics has been facilitated enormously by these methods. Techniques generally referred to as Markov Chain Monte Carlo (MCMC) have played a major role in this process,stimulating synergies among scientists in different fields, such as mathematicians, probabilists, statisticians, computer scientists and statistical geneticists.

1.1 Objectives of the Project

1. To make a survey of available literature on the topic
2. To study the possibilities of MCMC methods on genetic data
3. Develop R program to suit MCMC methods
4. Thus illustrate the MCMC sampling method, especially, Gibbs sampling procedure in the field of Genetics

2 Markov Chains

Applications of Markov Chains have been made to surprisingly diverse areas; learning theory, beginning with Estes (1950) and Bush and Mosteller (1955); information theory, Shannon (1948); changes in attitudes, Anderson (1954); labor and social mobility, Blumen, Kogen and Mc Caethy (1955), Berger (1957); epidemiology of mental disease, Marshall and Goldhammer (1955); brand loyalty and brand switching, Lipstein (1959) and Harrary and Lipstein

(1962). The availability of an excellent software package, BUGS, that uses Markov Chain Monte Carlo methods, makes Markov Chains favorable to Genetics as well (Thomas, Spiegelhater and Gilks 1992). Most applications have used only stationary Markov Chains.

A. A. Markov (1906) laid the foundations of the theory of finite Markov chains, but concrete applications remained confined largely to card shuffling and linguistic problems. The theoretical treatment was usually by algebraic methods. The theory of chains with infinitely many states was introduced by A. Kolmogorov. This new approach made the theory accessible to a wider public and drew attention to the variety of possible applications. Since then Markov chains have become a standard topic in probability and a familiar tool in many applications. The existence of an invariant measure for persistent chains was first proved by C. Derman(1954). Taboo probabilities as a powerful tool in the theory of Markov chains were introduced by Chung (1953). Time-reversed Markov chains were first considered by A. Kolmogorov. Exit and entrance boundaries were introduced by W. Feller.

2.1 Basic Concepts

A sequence of trials with possible outcome $E_1, E_2, \ldots$, is called a *Markov chain* if the probabilities of sample sequences are defined by $P[E_{j_0}, E_{j_1}, \ldots, E_{j_n}] = a_{j_0} p_{j_0 j_1} p_{j_1 j_2} \ldots p_{j_{n-2} j_{n-1}} p_{j_{n-1} j_n}$; in terms of a probability distribution $\{a_k\}$ for E_k at the initial (or zero-th) trial and fixed conditional probabilities p_{jk} of E_k given that E_j has occurred at the preceding trial. The possible outcome E_k are usually referred to as possible states of the system, instead of saying that the n^{th} trial results in E_k, one says that the n^{th} step leads to E_k. p_{jk} is called the probability of a transition from E_j to E_k. Also, it is assumed that the trials are performed at a uniform rate so that the number of the step serves as time parameter. The transition

probabilities p_{jk} will be arranged in a matrix form which is known as the transition probability matrix (t.p.m), $P = \begin{pmatrix} p_{11} & p_{12} & \dots \\ p_{21} & p_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix}$. P is a square matrix (finite or infinite) with non-negative elements and unit row sums. Such a matrix is called a stochastic matrix. Any stochastic matrix can serve as a t.p.m; together with the initial distribution $\{a_k\}$ it completely defines a Markov chain with states $E_1, E_2, \dots$.

$p_{jk}^{(n)}$ denote the probability of a transition from E_j to E_k in exactly n steps. And $p_{jk}^{(m+n)} = \sum_v p_{jv}^{(m)} p_{vk}^{(n)}$; a special case of *Chapman Kolmogorov identity*, where, for all $n \geq 0$, $p_{jj}^{(0)} = 1$; $p_{jk}^{(0)} = 0, \ \forall j \neq k$.

A state E_k can be *reached* from a state E_j if there exists some $n \geq 0$ such that $p_{jk}^{(n)} > 0$. A set C of states is *closed* if no state outside C can be reached from any state E_j in C. For an arbitrary set C of states the smallest closed set containing C is called the *closure* of C.

A single state E_k forming a closed set will be called *absorbing*. A Markov chain is *irreducible* if there exists no closed set other than the set of all states. Clearly, C is closed if and only if, $p_{jk} = 0$ whenever j is in C and k outside C. If in the matrices P^n all rows and all columns corresponding to states outside C are deleted, there remain stochastic matrices for which the fundamental relations.
$p_{jk}^{(n+1)} = \sum_v p_{jv} p_{vk}^{(n)}$ and
$p_{jk}^{(m+n)} = \sum_v p_{jv}^{(m)} p_{vk}^{(n)}$ again hold. That is, we have a Markov chain defined on C and this sub chain can be studied independently of all other states. The state of E_k is absorbing if and only if $p_{kk} = 1$ and a chain is irreducible if and

only if, every state can be reached from every other state. The state E_j has period $t > 1$ if $p_{jj}^{(n)} = 0$ unless $n = vt$ is a multiple of t, and t is the largest integer with this property. The state E_j is *aperiodic* if no such $t > 1$ exists. A state E_j to which no return is possible (for which $p_{jj}^{(n)} = 0$ for all $n > 0$) will be considered aperiodic.Let $f_{jk}^{(n)}$ denote the probability that in a process starting from E_j the first entry to E_k occurs at the n^{th} step. We have $f_{jk}^{(0)} = 0$ $f_{jk} = \sum_{n=1}^{\infty} f_{jk}^{(n)}$. That is, the probability that starting from E_j, the system will ever pass through E_k, $f_{jk} \leq 1$. When $f_{jk} = 1$, the $\{f_{jk}^{(n)}\}$ is a proper probability distribution and it is referred to as the first-passage distribution for E_k. In particular, $\{f_{jj}^{(n)}\}$ represents the distribution of recurrence times for E_j. When $f_{jj} = 1$, a return to E_j is certain.

$\mu_j = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$ is the mean recurrence time for E_j. State E_j is *persistent* if $f_{jj} = 1$ and *transient* if $f_{jj} < 1$ A persistent state is called *null state* if its mean recurrence time $\mu_j = \infty$ An aperiodic persistent state E_j with $\mu_j < \infty$ is called *ergodic*. In an irreducible chain with only ergodic elements the limits

$$u_k = lim_{n\to\infty} p_{jk}^{(n)}$$

exists and are independent of the initial state j.

A probability distribution $\{u_k\}$ satisfying $u_j = \sum_i u_i p_{ij}$ is called *invariant or stationary distribution*, for the given Markov chain. An irreducible aperiodic chain possesses an invariant probability distribution $\{u_k\}$ if and only if, it is ergodic. If a chain possesses an invariant probability distribution $\{u_k\}$ then $u_k = 0$ for each E_k that is either transient or a persistent null state. In ergodic chains, the probabilities $p_{jk}^{(n)}$ tend to the term u_k of the invariant probability distribution.

3 Monte Carlo Methods

Monte Carlo is the art of approximating an expectation by the sample mean of a function of simulated random variables. It is about invoking laws of large numbers to approximate expectations. Consider a (possibly multidimensional) rv X having distribution function $F_X(x)$ on a set of values $\mathcal{X}$. Then the expected value of a function g of X is,

$$E(g(X)) = \int_{\mathcal{X}} g(x) dF_X(x)$$

where integral is taken in Lebesque-Steiljes sense.

Now, if we were to take an n-sample of X's, $(x_1, x_2, \ldots, x_n)$, and we computed the mean of $g(x)$ over the sample, then, we would have the Monte Carlo estimate

$$\widetilde{g_n}(x) = \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$

of $E(g(X))$. We could alternatively speak of the rv

$$\widetilde{g_n}(X) = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

called the Monte Carlo estimator of $E(g(X))$. If $E(g(X))$ exists, then the weak law of large numbers tells us that for any arbitrarily small ε,

$$lim_{n\to\infty} P(\mid \widetilde{g_n}(x) - E(g(X)) \mid \geq \varepsilon) = 0.$$

Also,

$$\begin{aligned} E(\widetilde{g_n}(X)) &= E(\frac{1}{n} \sum_{i=1}^{n} g(X_i)) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(g(X_i)) \\ &= E(g(X)) \end{aligned}$$

that is, $\widetilde{g_n}(x)$ is unbiased for $E(g(X))$. This result becomes useful when one realizes that many quantities of interest may be cast as expectations.

The immediate consequence of this is that all probabilities, integrals, and summations can be approximated by the Monte Carlo method. And,

$$\begin{aligned} Var(\widetilde{g_n}(X)) &= Var(\frac{1}{n}\sum_{i=1}^{n} g(X_i)) \\ &= \frac{Var(g(X))}{n} \end{aligned}$$

An unbiased estimator for $Var(g(X))$ is

$$\begin{aligned} \widetilde{Var}(g(X)) &= \frac{1}{n-1}\sum_{i=1}^{n}(g(x_i) - \widetilde{g_n}(x))^2 \\ &= \sigma^2 \\ \widetilde{Var}(\widetilde{g_n}(X)) &= \frac{\widetilde{Var}(g(X))}{n} \\ &= \frac{\sigma^2}{n} \\ &= \int_{\mathcal{X}} [g(x) - E(g(X))]^2 dF_X(x) \end{aligned}$$

Hence as n increases, variance reduces. In general, the aims of Monte Carlo methods are to solve one or both of the following:

1.to generate samples $(x_1, x_2, \ldots, x_n)$ from a given pmf or pdf $f_X(x)$ known usually as the target density.

2.to estimate expectations of functions under this distribution.

The usual procedure is to generate samples from the target density $f^*(x) = f(x)/K$, where, $x \in \mathcal{R}^d$, f (x) is the unnormalized density, and K is the (possibly unknown) normalizing constant. Let $h(x)$ be a density (candidate or proposal density) that can be simulated by some known method, and suppose there is a known constant C such that $f(x) \leq C\ h(x)$ for all x. Then the following algorithm is executed to obtain a random variate from $f^*(.)$.

1. Generate a candidate Z from $h(.)$ and a value u from $U(0,1)$, the uniform distribution on $(0,1)$.

2. If $u \leq f(z)/C\ h(x)$, set $Z = y$.

3. Else go to step 1.

It is shown that the accepted value y is a random variate from $f^*(.)$ (S.Chib and E.Greenberg 1995). For this method to be efficient, C must be carefully selected.

4 MCMC Methods

Monte Carlo simulation methods simulate independent rvs identically distributed with density function $f_X(x)$, the target density. These methods work well if the proposal density $h(x)$ is similar to $f(x)$. In large and complex problems it is difficult to find a single density $h(x)$ that has this property. MCMC methods are a class of methods for sampling from probability distributions based on constructing a Markov chain that has the target distribution as its stationary(invariant) distribution. This involves a Markov process in which a sequence of states $\{x^t\}$ is generated, each sample x^t having a probability distribution that depends on the previous value x^{t-1}. The process is started at an arbitrary x and iterated a large number of times. After this large numbers, the distribution of the observations generated from the simulation is approximately the target distribution (invariant disrtibution of the chain). This convergence to the invariant distribution occurs under mild regularity conditions. The regularity conditions required are irreducibility and aperiodicity. The usual sampling methods are the Metropolis-Hastings(MH) Sampling and Gibbs Sampling.

4.1 Metropolis-Hastings Sampling

This method also makes use of a proposal density $q(x)$, which depends on the current state x^t to generate a new proposed sample x', $q(x' \mid x^t)$. It generates a Markov chain in which each state x^{t+1} depends only on the previous state x^t. As before, assume that we can evaluate $f^*(x)$ for any x. A tentative new state x' is generated from the proposal density $q(x' \mid x^t)$. Now, we compute the ratio

$$\alpha(x', x^t) = \frac{f^*(x')q(x^t \mid x')}{f^*(x^t)q(x' \mid x^t)}$$

If $\alpha \geq 1$, the new state is accepted, that is, $x^{t+1} = x'$, otherwise($\alpha < 1$),

$$x^{t+1} = \begin{cases} x' & \text{with probability } \alpha \\ x^t & \text{with probability } 1 - \alpha \end{cases}$$

When $\alpha < 1$, we generate a random variable u from uniform $U(0, 1)$ and if the value of $u < \alpha$, we accept x' as x^{t+1}, otherwise we set, $x^{t+1} = x^t$.
For any positive q, $(q(x' \mid x) > 0 \ \forall x \ , x')$, as $t \to \infty$, the probability distribution of X^t tends to $f(x)$.

4.2 Gibbs Sampling

To implement the M-H algorithm, it is necessary that a suitable proposal density be specified. This density is selected from a family of distributions that requires the specification of such tuning parameters as the location and scale.
One family of proposal densities is given by $q(x, y) = q_1(y - x)$, where $q_1(.)$ is a multivariate density. The candidate y is thus drawn according to the process $y = x + z$, where z is called the increment random variable and follows the distribution q_1. A second family of proposal densities is given by the form $q(x, y) = q_2(y)$. In contrast to the random walk chain, the candidates are drawn independently of the current location x. A third choice,

which seems to be an efficient solution when available, is to exploit the known form of $f^*(.)$ to specify a proposal density. A fourth method is to use the acceptance-rejection method with a pseudo dominating density. And a fifth family is represented by a vector autoregressive process of order 1.

Another method of applying M-H algorithm, called block-at-a-time or variable-at-a-time method simplifies the search for a suitable proposal density. Consider a two variable situation, $x = (x_1, x_2)$. Suppose that there exists a conditional transition kernel $P_1(x_1, dy_1 \mid x_2)$ with the property that, for a fixed value of x_2, $F^*_{1|2}(. \mid x_2)$ is its invariant distribution (with density $f^*_{1|2}(. \mid x_2)$). That is,

$$F^*_{1|2}(dy_1 \mid x_2) = \int P_1(x_1, dy_1 \mid x_2) f^*_{1|2}(x_1 \mid x_2) dx_1$$

Also suppose the existence of a conditional transition kernel $P_2(x_2, dy_2 \mid x_1)$ with the property that, for a fixed value of x_1, $F^*_{2|1}(. \mid x_1)$ is its invariant distribution (with density $f^*_{2|1}(. \mid x_1)$). $P(x, A)$ for $x \in R^d$, $A \in \mathcal{B}$ (the transition kernel), denote the conditional distribution function that represents the probability of moving from x to point in the set A. Now the product of the transition kernel has $f^*(x_1, x_2)$ as its invariant density (Product of Kernels principle as called by S.Chib and E.Greenberg (2003)). Letting

$$P_1(x_1, dy_1 \mid x_2) = F^*_{1|2}(dy_1 \mid x_2)$$

and

$$P_2(x_2, dy_2 \mid y_1) = F^*_{2|1}(dy_2 \mid y_1)$$

the resulting method is called Gibbs algorithm, in which case
$\alpha(x, y) = 1$ for all x, y.

Thus Gibbs sampling is a special case of the M-H method. Gibbs sampling is applicable in multivariate situations when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. Let $X = (X_1, X_2)$ with joint density $f_{X_1,X_2}(x_1, x_2)$. Here, we begin with a value x_2^0 and sample x_1^0 from the conditional density $f(x_1 \mid x_2^0)$. Next value

x_2^1 is sampled from $f(x_2 \mid x_1^0)$. Now x_1^1 is sampled from $f(x_1 \mid x_2^1)$, and so on. In general, if $X = (X_1, X_2, \ldots, X_k)$ with joint density $f_X(x)$, at the $(t+1)^{th}$ step, with $x^t = (x_1^t, x_2^t, \ldots, x_k^t)$,

$$\begin{array}{lcl} x_1^{t+1} & \sim & f(x_1 \mid x_2^t, x_3^t, \ldots, x_k^t) \\ x_2^{t+1} & \sim & f(x_2 \mid x_1^{t+1}, x_3^t, \ldots, x_k^t) \\ \ldots & \ldots & \ldots \\ x_k^{t+1} & \sim & f(x_1 \mid x_1^{t+1}, x_2^{t+1}, \ldots, x_{k-1}^{t+1}) \end{array}$$

and form, $x^{t+1} = (x_1^{t+1}, x_2^{t+1}, \ldots, x_k^{t+1})$. The collection of full conditional distributions uniquely determines the joint distribution, provided the joint distribution is proper. Gibbs sampling is a MH method where every proposal is always accepted and the probability distribution of X^t tends to $f_X(x)$ as $t \to \infty$, and hence, each component X_j^t is very nearly a random sample from the marginal distribution $f_{X_j}(x_j)$, for $j = 1, 2, \ldots, k$, provided t is sufficiently large.

4.3 Convergence of the Chain

The time series plot, obtained by plotting the random variable being generated against the number of iterations, is a method to study the behavior of the chain. A chain is said to be poorly mixing if it stays in small regions of the parameter space for long periods of time, as opposed to a well mixing chain that explore the whole space uniformly (Walley 1991).

5 Bayes' Theorem in Statistical Decision Theory

Statistical decision theory is concerned with the making of decisions in the presence of statistical knowledge which sheds light on some of the uncertainties involved in the decision problem. Classical statistics is directed towards the use of sample information in making inferences about θ. In decision

theory, an attempt is made to combine the sample information with other relevant aspects of the problem in order to make the best decision. In addition to the sample information, two other types of information are typically relevant. The first is a knowledge of the possible consequences of the decisions. Often this knowledge can be quantified by determining the loss that would be incurred for each possible values of θ. The second source of non sample information that is useful to consider is called prior information. This is information about θ arising from sources other than current statistical investigation. Generally, prior information comes from past experience about similar situations involving similar θ. To evaluate the probability of a hypothesis, the Bayesian probabilist specifies some prior probability, which is then updated in the light of new relevant data.

One reason why interest in Bayesian methods has flourished is because of the great strides in Bayesian computing. The fundamental work of Geman and Geman(1984) influenced Gelfand and Smith(1990) to write a new paper that sparked new interest in Bayesian methods, statistical computing, algorithms and stochastic processes through the use of MCMC methods such as the Gibbs sampler and the Metropolis-Hastings algorithm.

5.1 Bayes' Theorem

Bayes Theorem is an essential element of the Bayesian approach to statistical inference. Bayes Theorem is also referred to in the literature as the 'Principle of inverse probability'. *Bayes Theorem:*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, P(B) > 0,$$

where A and B are any two events associated with a random experiment.

Generalized Bayes Theorem: Let $A_1, A_2, \ldots,$ be an finite or infinite

sequence of mutually exclusive events with $\bigcup_{i=1}^{\infty} A_i = \Omega$, the sample space and $P(A_i) > 0$ for $i = 1, 2, \ldots$. Suppose $B \subset \Omega$ is any other event such that $P(B) > 0$, then,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j)P(A_j)}, \quad i = 1, 2, \ldots.$$

5.2 Bayesian Estimation

To estimate θ, a number of classical techniques can be applied to the posterior distribution. The most common classical technique is maximum likelihood estimation, which chooses, as the estimate of θ the value $\hat{\theta}$ which maximizes the likelihood function. The analogous Bayesian estimate is defined as follows.

Definition 5.1. The generalized maximum likelihood estimate of θ is the largest mode $\hat{\theta}$, of the posterior distribution $g(\theta \mid x)$.

An important element of many Bayesian Analysis is the prior information concerning θ. A convenient way to quantify such information is in terms of a probability distribution on $\Theta, g(\theta)$. There are different methods for prior density determination. The two most important and most sought after methods are conjugate method and method of noninformative priors.

Conjugate Families: Let $\mathcal{F}$ denote the class of density functions $f(x \mid \theta)$, indexed by θ. A class $\mathcal{P}$ of prior distributions is said to be conjugate family for $\mathcal{F}$ if $g(\theta \mid x)$, the posterior density, is in the class $\mathcal{P}$, for all $f \in \mathcal{F}$ and $g \in \mathcal{P}$. For a given class of densities $\mathcal{F}$, a conjugate family can frequently be determined by examining the likelihood functions $l_k(\theta) = f(x \mid \theta)$, and choosing as a conjugate family, the class of distributions with the same functional

form as these likelihood functions. When dealing with conjugate priors, there is generally no need to explicitly calculate $m(x)$, the unconditional marginal density of sample X. The reason is that since $g(\theta \mid x) = \frac{h(x,\theta)}{m(x)}$, the factors involving θ in $g(\theta \mid x)$ must be same as the factors involving θ in $h(x, \theta)$. Hence it is only necessary to look at the factors involving θ in $h(x, \theta)$, and if these can be recognized as belonging to a particular distribution. If so, $g(\theta \mid x)$ is that distribution. The marginal density $m(x)$ can then be determined, if desired, by dividing $h(x, \theta)$ by $g(\theta \mid x)$.

Non informative priors: Once a loss and prior have been chosen, the calculation of the Bayes rule will be seen to be straight forward. This makes appealing attempts to use the Bayesian approach even when no, or limited, prior information is available. What is needed in such situations is a non informative prior, by which is meant a prior which favours no possible values of θ over others (that is, which contains no information about θ). The simplest situation to consider is when Θ is a finite set, consisting of say, n elements. The obvious non informative prior is to then give each element of Θ probability $\frac{1}{n}$. Here, equality of chances is blended with our ignorance.

5.3 MCMC in Bayesian Analysis

Bayesian analysis is performed by combining the prior information $g(\theta)$ and the sample information x into the posterior distribution of θ given x, $g(\theta \mid x)$ from which all decisions and inferences about θ are made.

$$g(\theta \mid x) = \frac{h(x, \theta)}{m(x)}$$
$$= \frac{joint\ density\ of\ X\ and\ \theta}{marginal\ density\ of\ X}$$

$$= \frac{g(\theta)f(x \mid \theta)}{\sum_{\Theta} g(\theta)f(x \mid \theta)}$$

or

$$= \frac{g(\theta)f(x \mid \theta)}{\int_{\Theta} g(\theta)f(x \mid \theta)}$$

When the parameter space Θ is not one dimensional, Bayesian analysis involves evaluation of multiple integration or summation, and so, have been of limited use for many practical applications. MCMC procedures allow us to avoid these difficulties by simulating correlated sequences, or, first order Markov chains, such that the posterior density(target density here) $g(\theta \mid x)$ is the invariant distribution of the chain. Features of the posterior distributions (mean,median,mode,quantiles,etc.) are approximated by the corresponding features of sampled values.

6 Application of MCMC Method to Breast Cancer Study

Numerous studies have investigated the genetic transmission of breast cancer. Claus *et al*(1991) uses known genetic methods like segregation analysis to investigate the familial risk of breast cancer based on a large case-control study and concludes that a small number of affected cases were due to the presence of a rare autosomal dominant allele, where as a larger number of cases reported were non genetic.

Segregation analysis and goodness of fit tests of genetic models provided evidence for the existence of a rare dominant allele (A) leading to increased susceptibility to breast cancer. The effect of genotype on the risk of breast

cancer is shown to be a function of a woman's age. The life time risk of breast cancer for carriers of the abnormal allele (A) was estimated to be nearly 100%. Thus, persons with both the genotypes AA and Aa will be affected by the disease at some time during their life period.

In the present study, we propose *Gibbs sampling* from *Multinomial-Dirichlet distributions* to estimate the proportion of persons affected with breast cancer at different age groups.

6.1 Theoretical Basis

The variable $X = (X_1, X_2, \ldots, X_p)$ has the *Multinomial distribution* $(N; \theta_1, \theta_2, \ldots, \theta_p)$, where $\theta_i > 0,\ for\ i = i, 2, \ldots p \sum_{i=1}^{p} \theta_i = 1, \sum_{i=1}^{p} X_i = N$ and

$$P(X \mid N; \theta_1, \theta_2, \ldots, \theta_p) = \begin{pmatrix} N \\ x_1 x_2 \ldots x_p \end{pmatrix} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_p^{x_p}.$$

The marginal distribution of each X_i being

$X_i \sim Binomial(N, \theta_i)$;

$0 \leq \theta_i \leq 1,\ i = 1, 2, \ldots, p.$

The conjugate prior of the *Multinomial distribution* is the *Dirichlet distribution*, the multivariate generalization of beta distribution. Hence the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ has a prior distribution given by,

$\theta \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_p)$ whose density function is given by

$$g(\theta \mid \alpha_1, \alpha_2, \ldots, \alpha_p) = \frac{\Gamma(\Sigma \alpha_i)}{\prod(\Gamma \alpha_i)} \theta_1^{(\alpha_1 - 1)} \theta_2^{(\alpha_2 - 1)} \ldots \theta_p^{(\alpha_p - 1)}, \alpha_i > 0$$

and $0 \leq \theta_i \leq 1, \sum \theta_i = 1.$

Marginally, $\theta_i \sim Beta(\alpha_i, \sum_{k \neq i} \alpha_k)$, $i = 1, 2, \ldots, p$.
The posterior distribution of θ given X is,
$\theta \mid x \sim Dirichlet(x_1 + \alpha_1, x_2 + \alpha_2, \ldots, x_p + \alpha_p)$.

6.2 Study

Claus *et.al* (1991) estimated the proportions in the population affected by breast cancer at different age groups as given in the table 1:

Age (years)	Proportion
20 - 29	0.0167
30 - 39	0.1277
40 - 49	0.2314
50 - 59	0.1719
60 - 69	0.1266
70 - 79	0.2709
80 +	0.0548
Total	1.0000

Table 1: *Proportions of breast cancer patients*

Because of the extremely low occurrence of breast cancer in women before the age of 20 years, the probability of becoming affected with breast cancer before 20 years is assumed to be zero. In the present study, we considered the seven age groups as seven classes of a *multinomial distribution* (see table 1).

$$X_i^{(r)} = \begin{cases} 1 & \text{if } r^{th} \text{ person has breast cancer at an age belonging to agegroup } i \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, 7$; $r = 1, 2, \ldots, N$.
$\theta_i = P(X_i^{(r)} = 1 \mid X_{i-1}^{(r)} = 0); i = 1, 2, \ldots, 7$, with $X_0^{(r)} = 0$.

Different age groups are the ages at onset of the disease and age group 80^+ can be thought of as being not affected and hence, $\theta_7 = 1 - \sum_{i=1}^{6} \theta_i$.
Also, each $X_i^{(r)} \sim Bernoulli(\theta_i); i = 1, 2, \ldots, 7$, and,
$\theta_i = P$(occurrence of breast cancer in the age group i) $= E(X_i^{(r)})$.
Also, $X_i = \sum_r X_i^{(r)} \sim Binomial(N, \theta_i); i = 1, 2, \ldots, 7$.
Hence, $X = (X_1, X_2, \ldots, X_7) \sim Multinomial(N, \theta_1, \theta_2, \ldots, \theta_7)$.

A lot of non genetic parameters may lead to the occurrence of the disease to a non carrier (a person with two normal allele of the gene - aa); for example, the occurrence of ovarian cancer, late marriage, infertility etc. It is well known that the beta (dirichlet) is the conjugate prior of the binomial (multinomial). One of the many applications of the beta distribution in Statistics is in the context of bioassay. The most common use of the beta in bioassay is in modeling parameter of a binomial distribution. A typical application is in quantal bioassay, where 'success' may constitute detection of a tumor of a certain type in a certain organ. In more general settings, such as multinomial response vector, the multivariate generalization of the beta distribution, the Dirichlet, is often used as a model for the response vector; in this case the beta will appear as a model for the marginal distributions.
Hence, we let, $\theta_i \sim Beta(\alpha_i, \beta_i); i = 1, 2, \ldots, 7$, where $\beta_i = \sum_{k \neq i} \alpha_k$, and
$\theta = (\theta_1, \theta_2, \ldots, \theta_7) \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_7)$. Hence the prior distribution of θ is
$g(\theta) = Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_7)$. The posterior distribution $g(\theta \mid x)$ is
Dirichlet $(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_7 + x_7)$.

We begin Gibbs sampling by assigning to $\theta = (\theta_1, \theta_2, \ldots, \theta_7)$, the prior probabilities obtained from Table 1. A random sample is drawn from $Multinomial(N; \theta_1, \theta_2, \ldots, \theta_7)$, with $N = 100$. These values are taken as the

first parameter values of beta distributions for the initial values of the iterations. At the end of the iterations, we have a sample of size m, the number of simulations, from the posterior distribution, arrayed as $\begin{pmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_7^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_7^{(2)} \\ \dots & \dots & \dots & \dots \\ \theta_1^{(m)} & \theta_2^{(m)} & \dots & \theta_7^{(m)} \end{pmatrix}$. Each row is a sample from the joint posterior distribution and columns are samples from the marginal distributions. First column gives samples from $g(\theta_1 \mid x)$ and second column gives samples from $g(\theta_2 \mid x)$ and so on. Now, discarding the initial values and averaging over the sample size, we get the improved estimates of the parameters. The R program code used for estimation is given in Appendix I. The proportions are estimated in four runs of the program and the output for the program execution is given in the Table 2. The variances of the proportions for various runs of the program is given in the Table 3. We estimate the proportions using uniform prior and the estimated values for four runs of the program is given in the Table 4.The prior proportions and the estimated proportions are plotted in Figures 1, 2, 3 and 4. The convergence of the chains are depicted in Figures 5, 6 7 and 8.

6.3 Output Analysis

Table 2 depicts the output analysis of the study for the seven groups for four runs of the program in Appendix.

Invoking Table 2 and Figures 1 and 2, the value of $\widehat{\theta_3}$ and $\widehat{\theta_6}$ are larger, compared to the values of all other parameters, leading to the conclusion that, the age groups 40^+ and 70^+ are the highest probable ages at onset of breast cancer.

As is evident from Table 3, the extremely small variances of the iterated values illustrate the suitability of Gibbs sampling in the situation. Also, it

Age (years)	$\widehat{\theta}$	Run1	Run2	Run3	Run4
20 - 29	$\widehat{\theta_1}$	0.01007200	0.01033389	0.01965272	0.01007097
30 - 39	$\widehat{\theta_2}$	0.14942026	0.13028229	0.09966381	0.09102453
40 - 49	$\widehat{\theta_3}$	0.25005136	0.24093310	0.27035829	0.26979353
50 - 59	$\widehat{\theta_4}$	0.15055422	0.18986951	0.19958647	0.22040184
60 - 69	$\widehat{\theta_5}$	0.14980235	0.11028538	0.08959871	0.12051873
70 - 79	$\widehat{\theta_6}$	0.25036396	0.24954914	0.26031505	0.23963357
80 +	$\widehat{\theta_7}$	0.03999377	0.07052424	0.06031410	0.04930833
Total	-	1.00025792	1.001778	0.9994892	1.000751

Table 2: *MCMC estimates of Proportions*

Variances	Run1	Run2	Run3	Run4
$\widehat{\theta_1}$	0.000106198	0.0001073495	0.0001842645	0.0001019053
$\widehat{\theta_2}$	0.0012585498	0.001106535	0.0008381735	0.0008451778
$\widehat{\theta_3}$	0.0019129253	0.0017651658	0.0019923381	0.0019049895
$\widehat{\theta_4}$	0.0012695066	0.0015165971	0.0015674542	0.0017302377
$\widehat{\theta_5}$	0.0012813251	0.0009974279	0.0008056178	0.001045555
$\widehat{\theta_6}$	0.0018793944	0.0018086309	0.0019176001	0.0018403571
$\widehat{\theta_7}$	0.0003776579	0.0006548418	0.0005499078	0.0004634782

Table 3: MCMC estimates of Variances of $\widehat{\theta_i}$

is evident from the Figures 5 and 6, that, the chain that is generated is very efficient. The time plots of the generated chains are depicted in Figures 5 and 6. It is seen that all the chains mix well and converge quickly to stationary distributions, as there are no horizontal regions visible in the graphs.

We may discard the initial values of iteration, and use improper priors like $Uniform(0, 1)$ to get the initial values. The corresponding R code is $\mathbf{X[1,]} < -\mathbf{runif(7)}$. And the output of various runs is given in Table 4. The

use of noninformative prior also gives us the same results. This is evident from Table 4 and Figures 3, 4, 7 and 8.

$\widehat{\theta}$	Run1	Run2	Run3	Run4
$\widehat{\theta_1}$	0.01009671	0.01981130	0.01014169	0.0099534367
$\widehat{\theta_2}$	0.12982770	0.14016432	0.07934393	0.159175893
$\widehat{\theta_3}$	0.26999674	0.27926111	0.25974321	0.229998541
$\widehat{\theta_4}$	0.12967862	0.18090535	0.22011336	0.188592595
$\widehat{\theta_5}$	0.11975033	0.08004237	0.11018176	0.099838067
$\widehat{\theta_6}$	0.26946581	0.2497542	0.27010255	0.250795979
$\widehat{\theta_7}$	0.06970986	0.04947005	0.04989858	0.0060248363
Total	0.99862209	0.9994087	0.9995251	0.998603

Table 4: *Proportions using initial values from Uniform prior*

7 Conclusion

Present study agrees with the conclusion in Claus et. al (1991), that, there is higher probabilities for the age groups 40^+ and 70^+ to be the ages at onset of disease, for those getting affected. MCMC analysis proves to be highly effective in estimating proportions of patients with a particular disease in a selected population.

It is usually difficult to localize genes that cause diseases with late ages at onset. These diseases frequently exhibit complex modes of inheritance, and only recent generations are available to be genotyped and phenotyped. In such situations, multipoint analysis using traditional exact linkage analysis methods, with many markers and full pedigree information, is a computationally intractable problem. MCMC sampling provides a tool to address this issue. This is effective not only in cancer studies but in the study of diseases like Alzheimer.

7.1 Publication

(i) Application of Multinomial-Dirichlet Cojugate in MCMC Estimation: A Breast cancer Study- published in *International Journal of Mathematical Analysis*, vol.4,(2010), no.41, pp 2043-2049.

7.2 Paper Presentation

(i) Presented a paper entitled *Applications of Bayesian Theory in Genetics* at the National Seminar held at NAS College, Kanhangad, during 12-14, Feb. 2009.

Bibliography

1. Andersen,S.K.,Olesen,K.G.,Jensen,F.V., Jensen,F.,(1989), HUGIN-a shell for Building Belief Universes for Expert Systems. *In Proc. 11th International Joint Conference on Artificial Intelligence,*1080-1085.

2. Beerenwinkel,N.,Antal,T.,Dingli,D., Traulsen,A.,K inzler,K.W., Velculescu,V.E.,Vogelstein,B., Nowak,M.A.,(2007), Genetic Progression and the Waiting Time to Cancer, *PLoS Comp. Bio.,*3(11),2239-2246.

3. Beerenwinkel,N.,Rahnenfuhrer,J.,Daumer,M., Hoffman,D.,Kaiser,R., Selbig,J.,
   Lengauer,T., (2005),Learning Multiple Evolutionary Pathways from Cross-Sectional Data,*Journal of Computational Biology*, 12(6),584-598.

4. Berger, J.O.,(1980),*Statistical Decision Theory*,Springer Verlag, N. York.

5. Cannings,C.,Thompson,E.A.,Skolnik, M. H.,(1978),Probability Functions on Complex Pedigrees, *Adv. in Appl. Probab.* ,10,26-61.

6. Carlin,B.P.,Louis,T.A.,(2000), *Bayes and Empirical Bayes methods for data analysis*,Chapman & Hall , CRC Press, Boca Raton.

7. Chib, S.,(2001), Markov Chain Monte Carlo Methods : Computation and Inference, *Handbook of Econometrics*, vol. 5 (eds. J J Heckerman and E Leamer), Elsevier, Amsterdam, 3569 - 3649.

8. Chib, S.,Greenberg, E.,(1995), Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, vol. 49,4,327-335.

9. Claus, E.B.,Risch,N.,Thompson,W.D.,(1990a),Age of Onset as an Indicator of Familial Risk of Breast Cancer,*Am. J. Epidemiol.,*131,961 - 972.

10. Claus,E.B.,Risch,N.,Thompson,W.D., (1990b),Using Age of Onset to Distinguish Between Sub forms of Breast Cancer,*Ann. Hum. Genet.,*54,169 - 177.

11. Claus,E.B.,Risch,N.,Thompson,W.D., (1991),Genetic Analysis of Breast Cancer in the Cancer and Steroid Hormone Study,*Am. J. Hum. Genet.,* 48,
232 - 242.

12. Cottingham,R.W.,Idury,R.M.,Schaffer, A.A.,(1993),Faster Sequential Genetic Linkage Computations,*Amer. J. Human Genetics,*53, 252-263.

13. Dawid,A.P.,(1979a),Conditional Independence in Statistical Theory,*J. Roy. Statist. Soc.,*B41,1-31.

14. Dawid,A.P.,(1979b),Some Misleading Arguments Involving Conditional Independence,*J. Roy. Statist. Soc.*,B41,249-52.

15. Dawid,A.P.,(1980a),Conditional Independence for Statistical Operations, *Ann. Statist.,*8,598-617.

16. Dojer,N.,Gambin,A.,Wilczynski,B.,Tiuryn,J.,(2006), Applying Dynamic Bayesian Networks to Perturbed Gene Expression Data,*BMC Bioinformatics,*7,249-260.

17. Egeland,T.,Mostad,P.F.,Mevag,B., Stenersen,M.,(2000),Beyond Traditional Paternity and Identification Cases:Selecting the Most Probable Pedigree,
*Forensic Sci. Int.,*110,47-59. Efron,

18. Efron,B.,Tibshirani,R.,(1994), *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

19. Elston,R.C.,Stewart,J.,(1971). A General Model for the Genetic Analysis of Pedigree Data,*Human Heredity,*21,523-542.

20. Falconer,D.S.,Mackay,T.F.C.,(1996). Introduction to Quantitative Genetics, 4th ed. Addison Wesley Longman Limited, Harlow, UK.

21. Feller, (1971),*An Introduction to Probability Theory and its Applications, Vol 1,* John Wiley & Sons, Inc..

22. Gelman,A.J.,Carlin,B.,Stern H., Rubin, D.,(2004),*Bayesian Data Analysis,*Chapman and Hall, London, 2^{nd} ed.

23. Geweke,J.,(1999),Using Simulation Methods for Bayesian Econometric Models : Inference, Development, and Communication ,*Econometric Reviews,*8,1 - 126.

24. Geweke,J.,(2005)*Contemporary Bayesian Econometrics and Statistics,*Wiley, NJ.

25. Ghahramani,Z.,(1997),Learning Dynamic Bayesian Networks,*Lecture Notes in Computer Science,*1387,168-197.

26. Gilks,W.R.,Gelman,A.,Roberts,G.O.,(1997),weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms, *The Annals of Appl. Prob.,*7(1),
110-120.

27. Go,R.C.P.,King,M.C.,Bailey-Wilson,J., Elston,R.C.,Lynch,H.T.,(1983), Genetic Epidemiology of Breast Cancer and Associated Cancers in High-Risk Families.I.Segregation analysis,*J. Natl. Cancer Institute.,*71,455 - 461.

28. Golstein,A.M.,Haile,R.W.C., Hodge,S. E.,Paganini-Hill,A.,Spence,M. A.,(1988), Possible Heterogeneity in the Segregation Pattern of Breast Cancer in Families with Bilateral Breast Cancer, *Genet. Epidemiol.,*5,121 - 133.

29. Golstein,A.M.,Haile,R.W.C.,Marazita,M. L.,Paganini-Hill,A.,(1987), A Genetic Epidemiologic Investigation of Breast Cancer in Families with Bilateral Breast Cancer.I.Segregation Analysis, *J.Natl.Cancer Institute.,*78,911 -918.

30. Greenberg,E.,(2003), *Introduction to Bayesian Econometrics*, Cambridge University Press.

31. Hitchcock,D.B.,(2003),A history of the Metropolis-Hastings Algorithm, *The American Statistician,*57(4),254-257.

32. Koop,G.,(2003),*Bayesian Econometrics*,Wiley,Chichester.

33. Koop,G.,Poirier,D.J.,Tobias,J. L.,(2003),*Bayesian Econometric Methods*, Cambridge University Press.

34. Mange,E.J.,Mange,A.P.,(1997), *Basic Human Genetics,*Rastogi Publications.

35. Meila,M.,Jordan,M.I.,(2001),Learning with Mixtures of Trees,*J. Mach Learn. Res.,*1,1-48.

36. Ramsey,F.L.,(1972),A Bayesian Approach to Bioassay,*Biometrics,* 28,841 - 858.

37. Rahnenfuhrer,J.,Beerenwinkel,N.,Schulz,W.A., Hartmann,C., Deimling,A.,Wullich,B.,Lengauer,T.,(2005),Estimating Cancer Survival and Clinical Outcome based on Genetic Tumor Progression Scores,*Bioinformatics,* 21,10,2438-2446.

38. Rizzo.M.L.,(2008),*Statistical Computing with R,*Chapman and Hall/CRC,New York.

39. Sattin,R.W.,Rubin,G.L.,Webster,L.A., Huezo,C.M.,Wingo,P.A., Ory,H.W.,Layde,D.M.,(1985), Cancer and Steroid Hormone Study: Family History and Risk of Breast Cancer,*JAMA,*253,1908 - 1913.

40. Smith,R.T.,(1986),Beta Distribution in Bioassay,*Handbook of Beta distribution and its applications,* Statistics : a Dekker series of Text Books and Monographs (eds. Gupta A K., Nadarajah S.), 437 - 455.

41. Tanner,M.A.,(2004),*Tools for Statistical Inference,*3^{rd} ed., Springer series in Statistics.

42. Thompson,E.A.,(2000) *Statistical Inference from Genetic Data on Pedigrees,*IMS,Beachwood,OH.

43. Varian,H.,(2005), Bootstrap Tutorial, *Mathematica Journal*, 9, 768-775.

44. Walley,P.,(1991),Inferences from Multinomial Data:Learning about a bag of marbles,*J.R.Statistic.Soc.B,*58(1),3-57.

45. Walsh,B.,(2004),Markov Chain Monte Carlo and Gibbs Sampling,*Lecture Notes for EEB 581.*

Appendix

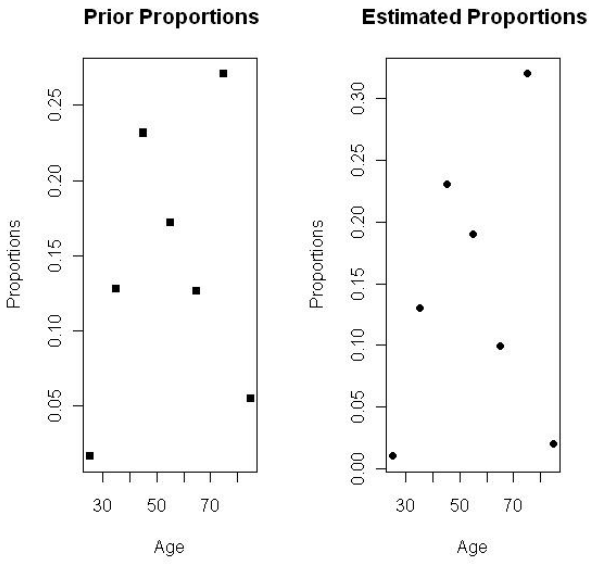

Figure 1: Proportions for Run I

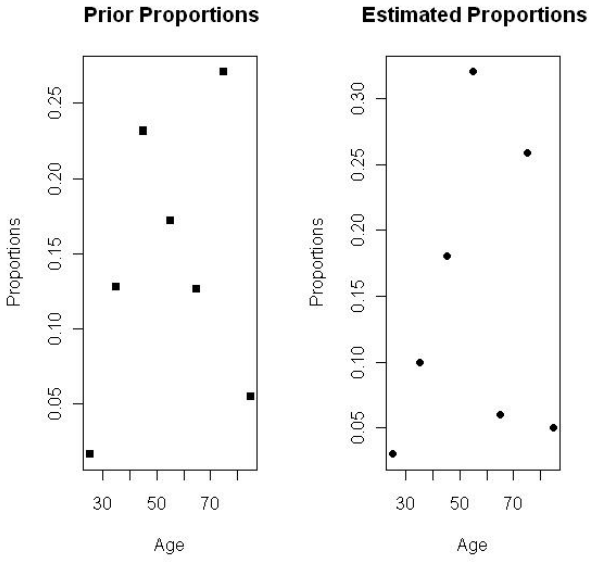

Figure 2: Proportions for Run II

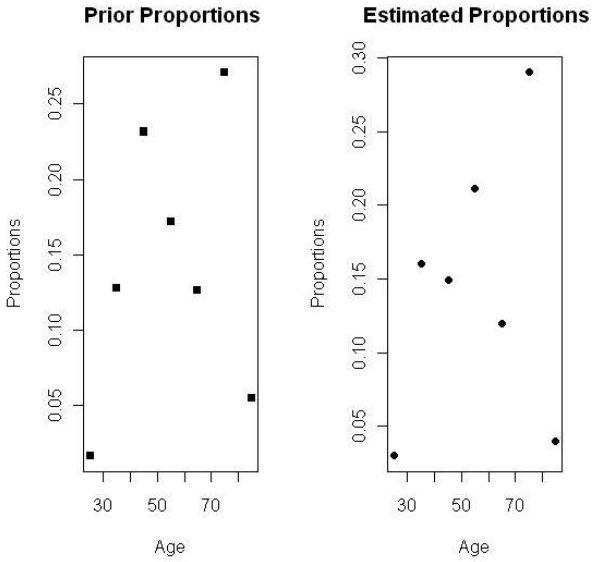

Figure 3: Proportions for Run I under Uniform prior

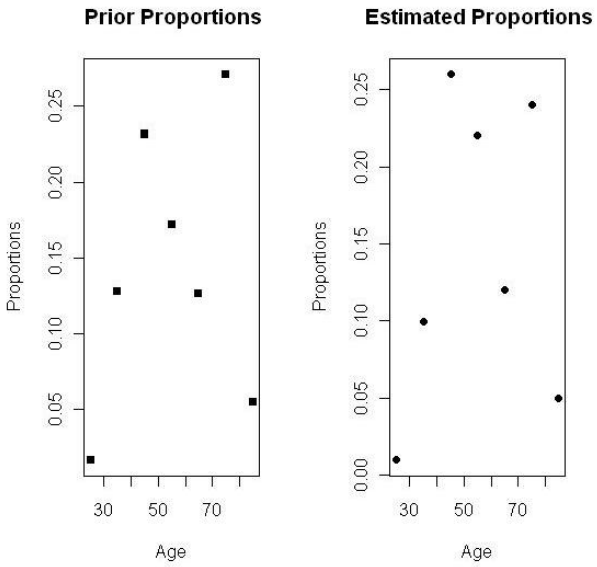

Figure 4: Proportions for Run II under Uniform prior

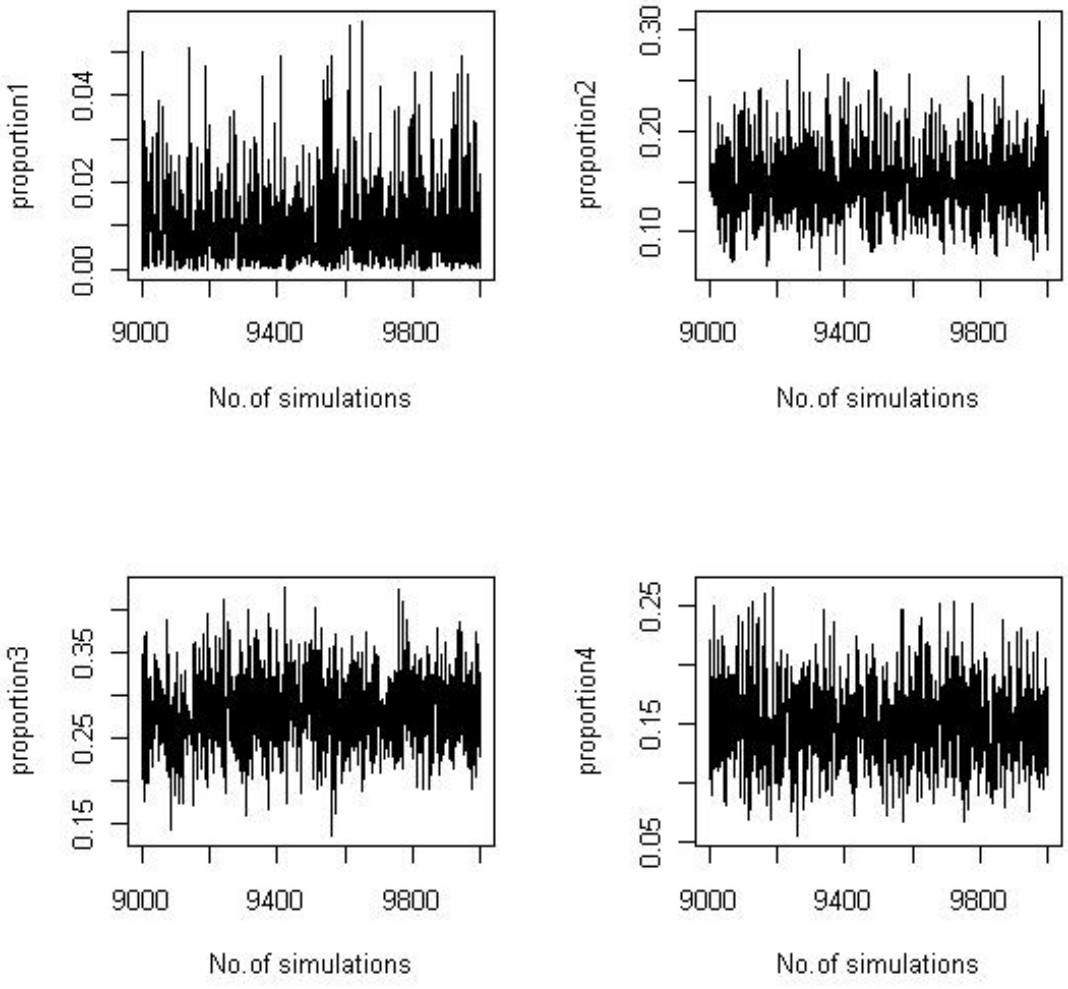

Figure 5: Convergence of the Chains

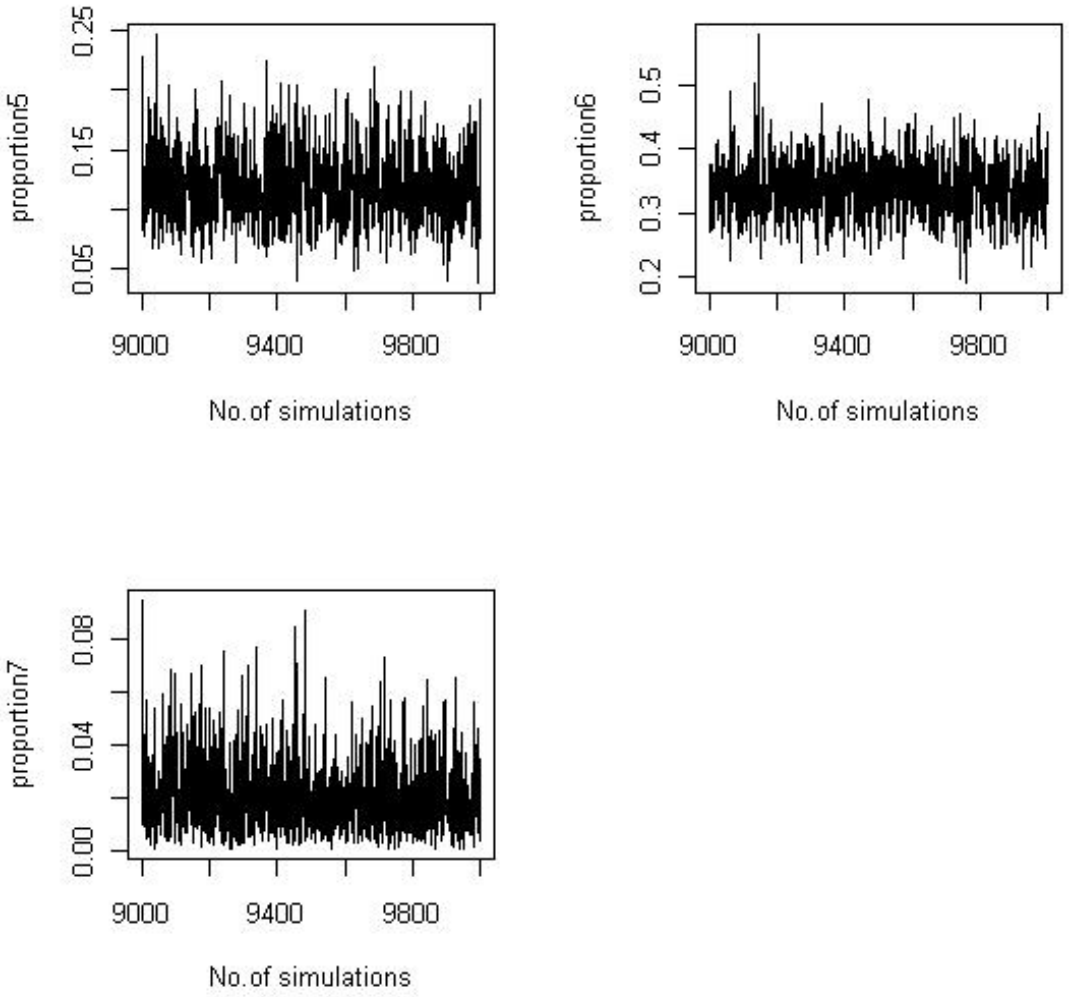

Figure 6: Convergence of the Chains

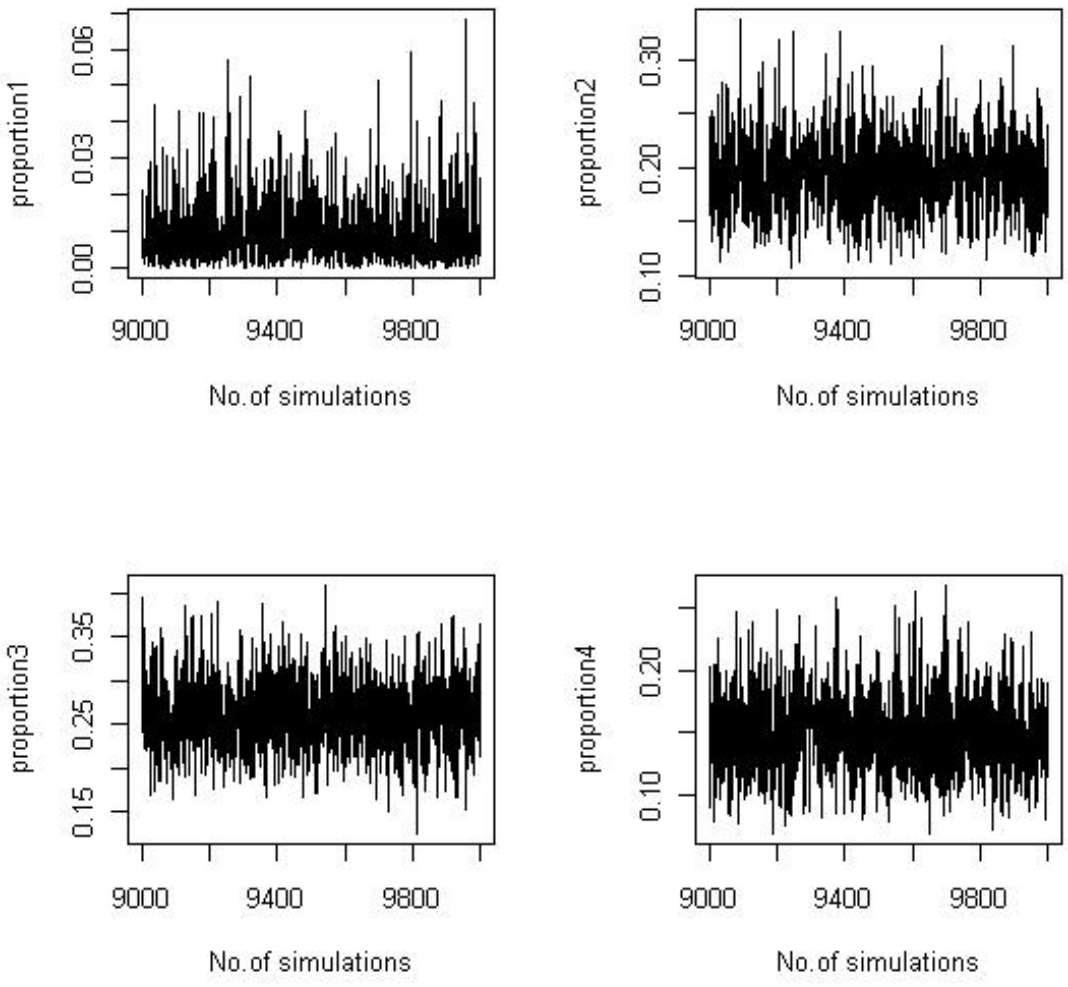

Figure 7: Convergence of the Chains under Uniform prior

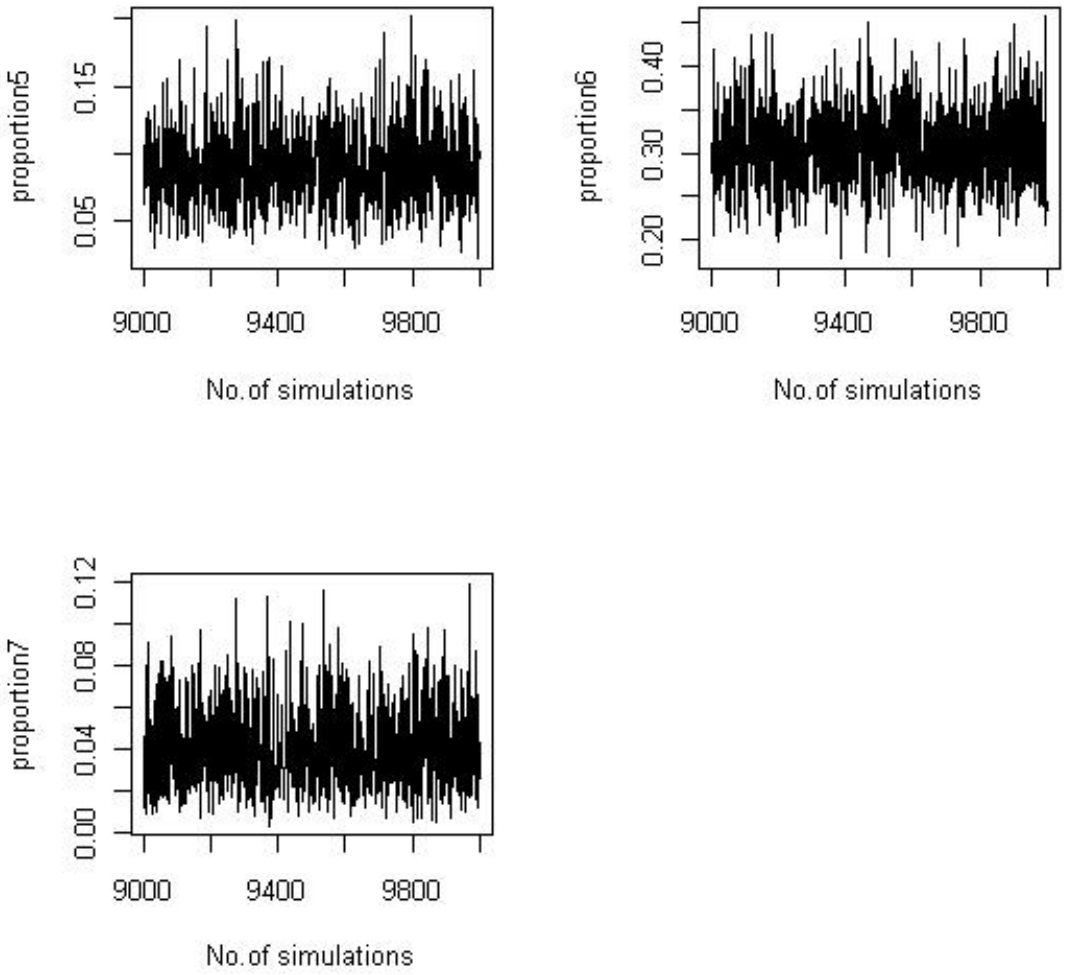

Figure 8: Convergence of the Chains under Uniform prior

```
#R code to estimate the parameters of multinomial
distribution using Gibbs sampling

# Assigning prior values to the parameters
N<-100
p<-
c(0.0167,0.1277,0.2314,0.1719,0.1266,0.2709,0.0548)

#Random sampling from multinomial distribution
M(N,p1,p2,p3,p4,p5,p6,p7)
s<-rmultinom(1,N,p)  #assigns the values of a single
random sample to the first column of the matrix s

#Gibbs sampling
m<-10000;
x<-matrix(0,m,7)     #x is a matrix of m rows,7 columns
and zero entries
g1<-s[1,1];g2<-s[2,1];g3<-s[3,1];g4<-s[4,1];
g5<-s[5,1];g6<-s[6,1];g7<-s[7,1];
 #Multinomial random sample values assigned to g1:g7

b1<-g2+g3+g4+g5+g6+g7
b2<-g1+g3+g4+g5+g6+g7
b3<-g1+g2+g4+g5+g6+g7
b4<-g1+g2+g3+g5+g6+g7



b5<-g1+g2+g3+g4+g6+g7
b6<-g1+g2+g3+g4+g5+g7
b7<-g1+g2+g3+g4+g5+g6

x[1,1]<-rbeta(1,g1,b1);x[1,2]<-rbeta(1,g2,b2);
x[1,3]<-rbeta(1,g3,b3);x[1,4]<-rbeta(1,g4,b4);
x[1,5]<-rbeta(1,g5,b5);x[1,6]<-rbeta(1,g6,b6);
x[1,7]<-rbeta(1,g7,b7)

x[1,]        #initial values as the first row of the matrix x
```

```
for(i in 2:m)
{
x[i,1]<-rbeta(1,g1,b1)
x[i,2]<-rbeta(1,g2,b2)
x[i,3]<-rbeta(1,g3,b3)
x[i,4]<-rbeta(1,g4,b4)
x[i,5]<-rbeta(1,g5,b5)
x[i,6]<-rbeta(1,g6,b6)
x[i,7]<-rbeta(1,g7,b7)
}
x1<-mean(x[5000:10000,1])
x2<-mean(x[5000:10000,2])
x3<-mean(x[5000:10000,3])
x4<-mean(x[5000:10000,4])
x5<-mean(x[5000:10000,5])
x6<-mean(x[5000:10000,6])
x7<-mean(x[5000:10000,7])
```

```
Estimate<-c(x1,x2,x3,x4,x5,x6,x7);Estimate
Sum<-Estimate[1]+ Estimate[2]+ Estimate[3]+
Estimate[4]+ Estimate[5]+ Estimate[6]+ Estimate[7];
Sum
s1<-var(x[5000:10000,1]); s2<-var(x[5000:10000,2]);
s3<-var(x[5000:10000,3]);s4<-var(x[5000:10000,4]);
s5<-var(x[5000:10000,5]); s6<-var(x[5000:10000,6]);
s7<-var(x[5000:10000,7]);
Var<-c(s1,s2,s3,s4,s5,s6,s7);Var;
par(mfrow=c(1,2));
u<-seq(25,85,10);
plot(u,p,main="PriorProportions",
xlab="Age",ylab="Proportions",pch=15,type="p");
plot(u,E,main="EstimatedProportions",xlab="Age",y
lab="Proportions",
pch=16,type="p");
```

```
par(ask=TRUE);
par(mfrow=c(3,2));
index<-5000:5500
y1<-x[index,1];y2<-x[index,2];y3<-x[index,3];
y4<-x[index,4];y5<-x[index,5];y6<-x[index,6]
plot(index,y1,type="l",main="",ylab="proportion1");
plot(index,y2,type="l",main="",ylab="proportion2")
plot(index,y3,type="l",main="",ylab="proportion3");
plot(index,y4,type="l",main="",ylab="proportion4")
plot(index,y5,type="l",main="",ylab="proportion5");
plot(index,y6,type="l",main="",ylab="proportion6")
par(ask=TRUE);
index1<-9000:10000;y7<-x[index1,7]
par(mfrow=c(1,1));
plot(index1,y7,type="l",main="",ylab="proportion7"
)
```