CHAPTER I

A REVIEW OF MARKOV CHAIN MONTE CARLO METHODS
CHAPTER I

A REVIEW OF MARKOV CHAIN MONTE CARLO METHODS

1.0 INTRODUCTION

Markov Chain Monte Carlo (MCMC) is a powerful technique for performing integration by simulation. In recent years MCMC has revolutionized the application of Bayesian statistics. Many high dimensional complex models, which were formally intractable, can now be handled routinely. MCMC has also been used in specialized non-Bayesian problems. A good introduction on MCMC methods in bio statistical applications can be found in Gilks et al (1996) and Gelman and Rubin (1996). The techniques have been applied in most areas of statistics and Biostatistics namely vaccine efficacy, genomics, proteomics, clinical monitoring, pharmacokinetics, disease mapping, image analysis, genetics and epidemiological research. Gilks et al (1996), Berry and Stangle (1996) describe applications in decision analysis, clinical trial design, and cross over trials, meta analysis, change point analysis, hemodynamics and prenatal mortality. The applications of MCMC in modeling situations involve hierarchical models, missing data, censored data, and spatially correlated data. MCMC methods have also been used extensively in statistical physics.
over the last 40 years, in spatial statistics for the past 20 years and in Bayesian image analysis over the last 10 years Gilks et al (1996). In the last 5 years MCMC has been introduced into significance testing, general Bayesian inference and maximum likelihood estimation.

The use of MCMC was first introduced in statistical mechanics by Metropolis et al (1953) to study the equation of the state of a two-dimensional rigid sphere system. The subject in its most general terms was presented by Fosdick (1963) and Binder (1992). The choice made by Metropolis et al (1953) was one of many other possibilities. The well-known Gibbs sampler as proposed by Geman and Geman (1984) and other options as given by Muller et al (1973) are also applicable to various problems. Metropolis et al (1953) introduced MCMC as a general method suitable for fast computing machines of calculating the properties of any substance considered as composed of interacting individual molecules. Now this method has become a miraculous tool of Bayesian analysis (Geyer, 1992) and the flag of what has been called as the model liberation movement by Smith (1992). Bayesian calculations not analytically tractable can be performed once a likelihood and prior are given (Besag et al, 1995). For non-Bayesian applications MCMC is considered as a very powerful numerical device in likelihood analysis or decision theory (Geyer, 1992).
Convergence is necessary for error estimation unless the system has reached stationarity both in averages and variance estimates may be meaningless. Different convergence diagnostics have been proposed and a far reaching discussion is provided by Cowles and Carlin (1996) who compared the performances of thirteen diagnostics. Green and Han (1992) reviewed a number of variants of estimators based on correlation functions and provided some comparisons concerning approximations and computational efforts. The Monte Carlo methods provide approximate solution to variety of mathematical problems by performing statistical sampling experiments on a computer. The method applies to problems with no probabilistic content as well as to those with inherent probabilistic structure. The method is called after the city in the Monaco principality, because of roulette, a simple random number generator. The name and the systematic development of Monte Carlo methods dates back about 1944.

Monte Carlo techniques are used in situation in which the analytic solution of the problem is either intractable or time consuming. Instead of calculating exact quantities, simulation is used to produce stochastic approximation to the solutions. Monte Carlo techniques have the long history in mathematics. Earliest example is Buffoon’s needle dating from 1733 A.D described in Morgan (1984). Realistically the wide spread application of Monte Carlo techniques has only become feasible with the availability of cheap and efficient
computing since the 1970's. In practice Monte Carlo methods are discussed interchangeably with simulation. Useful discussion can be found in a variety of monographs such as Morgan (1984), Ripley (1987), and Tanner (1993). Common uses are the construction of Monte Carlo hypothesis tests, bootstrap distributions, numerical integration and Bayesian calculations. More specialized uses are the analyses of complex stochastic systems. In fact the exponential increase in available computing power has led to Monte Carlo techniques facilitating major statistical advances. An important point is that the term Monte Carlo does not refer to a particular stochastic algorithm. The choice of algorithm depends on a variety of factors. In many problems there is a natural formulation. For example, we estimate the mean of the distribution by taking the mean of a sample from the distribution. Another criterion that should be considered when choosing an algorithm is the efficiency of the method. This leads to the contemplation of variance reduction techniques, which are used to produce more accurate estimates for the same computational effort.

In practice there is a trade off between the use of an inefficient algorithm that is easy to design and implement and an efficient algorithm that could take detailed analysis to design provided that the inefficient algorithm is not pathologically sufficient merely to run the algorithm for an appropriately longer period of time to obtain results comparable to those with the efficient algorithm. An important point is
that great care must be taken to ensure that satisfactory convergence has occurred. In any application the relevant literature should be consulted.

1.1 MARKOV CHAIN

Markov Chains are Markov processes whose time-dependent random variables (the state of the Markov chain) can assume values in a discrete set (the state space), either finite or countable infinite. The Markov property is essentially a conditional independence of the future evolution on the past (the whole history of the process being summarized in the current state). Basically, the chain can be seen as modeling the position of an object in a discrete set of possible locations over time, the next location being chosen at random from a distribution that depends only on the current one.

A Markov Chain is a stochastic process with the Markov property. In such a process, the distant past is irrelevant given knowledge of the recent past. In the discrete time case, the process consists of a sequence $X_1, X_2, ...$ of random variables. The domain of these variables is called the states space, with the value of $X_n$ being the state at time $n$. If the conditional distribution of $X_{n+1}$ on past states is a function of $X_n$,

$$P(X_{n+1}/ X_0, X_1, ..., X_n) = P (X_{n+1}/ X_n) \quad (1.1.1)$$
the process is said to have the Markov property. Markov Chains are named after A.A. Markov who produced the first result (1906) for these processes. A generalization to countably infinite state spaces was given by Kolmogorov (1936). Markov Chains are related to Brownian motion and the Ergodic hypothesis, two topics in physics which were important in the early years of the 20th century, but Markov appears to have been motivated by a purely mathematical approach, namely the extension of the law of large numbers to dependent events.

1.1.1 Properties of Markov Chain

A Markov Chain is characterized by the conditional distribution $P(X_{n+1} / X_n)$ which is called the transition probability of process. This is called one-step transition probability. The probability of a transition in 2, 3 or more steps are derived from one step transition probability and the Markov property.

$$P(X_{n+2} / X_n) = \int P(X_{n+2}, X_{n+1} / X_n) \, dX_{n+1}$$

$$= \int P(X_{n+2} / X_{n+1}) \, P(X_{n+1} / X_n) \, dX_{n+1} \quad (1.1.2)$$

Similarly,

$$P(X_{n+3} / X_n) = \int P(X_{n+3}/X_{n+2}) \, dX_{n+2}.$$  

$$\int P(X_{n+2} / X_{n+1}) \, P(X_{n+1} / X_n) \, dX_{n+1} \, dX_{n+2} \quad (1.1.3)$$

These formulas generalize to arbitrary future times $n+k$ by multiplying the transition probability and integrating $k$ times.
The marginal distribution \( P(X_n) \) is the distribution over the states at time \( n \). The initial distribution is \( P(X_0) \). The evolution of the process through one step is described by

\[
P(X_{n+1}) = \int P(X_{n+1} \mid X_n) P(X_n) \, dX_n
\]

(1.1.4)

This is a version of the Frobenius-Perron equation. There may exist one or more state distributions \( \pi \) such that

\[
\pi(X) = \int P(X \mid Y) \pi(Y) \, dY
\]

(1.1.5)

where \( Y \) is a convenient name for the variable of integration. Such a distribution \( \pi \) is called stationary or steady state distribution. A stationary distribution is an eigen function of the conditional distribution function, associated with the eigen value 1. Whether or not there is a stationary distribution and whether or not it is unique if it does exist, are determined by certain properties of the process. Irreducible means that every state is accessible from every other state. Aperiodic means there exist at least one state for which the transition from that state to itself is possible. Positive recurrent means that the expected return time is finite for every state. Some times the terms indecomposable, acyclic and recurrent are used as synonyms for irreducible, aperiodic and recurrent respectively. If the Markov Chain is the positive recurrent, there exists a stationary distribution. If it is positive recurrent and irreducible there exists a unique stationary
distribution and furthermore the process constructed by taking the stationary distribution as the initial distribution is ergodic. Then the average of a function $f$ over samples of the Markov Chain is equal to the average with respect to the stationary distribution,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(x_k) = \int f(X) \pi(X) dX$$  \hspace{1cm} (1.1.6)$$

In particular this holds for $f$ equal to the identity function. Thus the average of sample values over time is equal to the expected value of the stationary distribution. Also the equivalence of averages holds if $f$ is the indicator function on some subset $A$ of the state space.

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \pi_A(x_k) = \int_A \pi(X) dX = \mu_A(A)$$  \hspace{1cm} (1.1.7)$$

where $\mu_\pi$ is the measure induced by $\pi$. This makes it possible to approximate the stationary distribution by a histogram or other density estimate of a sequence of samples.

1.1.2 Discrete Markov Chain

If the state space is finite, the transition probability distribution can be represented as a matrix called the transition matrix, with $(i,j)^{th}$ element equal to $P(X_{n+1} = i | X_n = j)$. In this formulation, element $(i, j)^{th}$ is the probability of the transition from $j$ to $i$. For discrete state space, the integration in the $k$-step transition probability is summations
and can be computed as the $k^{th}$ power of the transition matrix. That is if $P$ is the one step transition matrix then $P^k$ is the transition matrix for the $k$-step transition. Using $P$ for the transition matrix, the stationary distribution is a vector, which satisfies the equation $\pi P = \pi$. In this case the stationary distribution is an eigen vector of the transition matrix associated with the eigen value 1. If the transition matrix $P$ is positive recurrent, irreducible and aperiodic then $P^k$ converges elementwise to a matrix in which each row is the unique stationary distribution.

A transition matrix which is positive is irreducible, aperiodic and positive recurrent. A matrix is a stochastic matrix iff it is a matrix of transition probabilities of some Markov Chain. Markov Chains are used to model various processes in queuing theory and statistics and can also be used as a signal model in entropy coding techniques. Markov Chains also have many biological applications, particularly population processes that are useful in modeling processes that are analogous to biological population. Markov Chains have been extensively used in bioinformatics applications. An example is genefinding algorithms for coding region / gene prediction.

Consider a sequence $\{ \xi_j \}_{j \in \mathbb{N}}$, where $\xi_j$ are i.i.d random variables such that $\xi_j = \pm 1$ with probability $1/2$, and define the sequence $\{ X_n \}_{n \in \mathbb{N}}$ by the following recurrence relation:
\[ X_0 = 0, \ X_n = X_{n-1} + \xi_n \mod N, \]

This defines a random walk on \( \{0,1,2, \ldots, N-1\} \). Clearly, given the present state \( X_n \), the distribution of \( X_{n+1} \) is determined, and one does not need to know about the past. This can be restated as the Markov property

\[ P(X_{n+1} = i_{n+1} / \{ X_m = i_m \}_{m=1}^n) = P(X_{n+1} = i_{n+1} / X_n = i_n) \quad (1.1.9) \]

i.e. the probability of \( X_{n+1} \) conditional on the whole past history of the sequence, \( \{X_m\}_{m=1}^n \), reduces to probability of \( X_{n+1} \) conditional on the latest value alone, \( X_n \). This last conditional probability is called the transition probability and it is of special importance for Markov chains. In this particular case, it is given by

\[ P(X_{n+1} = i \pm 1 / X_n = i) = P(\xi_{n+1} = \pm 1) = \frac{1}{2} \quad (1.1.10) \]

\[ P(X_{n+1} = \text{anything else} / X_n = i) = 0 \quad (1.1.11) \]

A Markov Chain does not need to be given as explicitly as in the recurrence relation above. More generally, given a state space \( S \) which we will assume finite for simplicity, one can define the transition probabilities, i.e. the probabilities that the process is in state \( i \) at time \( n+1 \) given that it was in state \( j \) at time \( n \)

\[ p_n(i / j) = P(X_{n+1} = i / X_n = j) \quad (1.1.12) \]
The Markov Chain is called stationary if $p_n(i/j)$ is independent of $n$, and from now on we will discuss only stationary Markov Chains and let $p(i/j) = p_n(i/j)$. The Markov property then requires that

$$P(X_{n+1} = j/X_n = k_n, X_{n-1} = k_{n-1} \ldots X_1 = k_1) = P(X_{n+1} = j/X_n = k_n) \quad (1.1.13)$$

i.e., the value of $X_{n+1}$ is determined completely by $X_n$ and does not depend on the past history of $X_n$. This property allow to express the probability of any finite sequence $(X_0, X_1, \ldots, X_n)$ in terms of the initial distribution $\mu(i) = P(X_0 = i)$ and $p(i/j)$. Indeed, the Markov property implies that

$$P(X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p(i_n/i_{n-1})p(i_{n-1}/ i_{n-2}) \ldots p(i_1/i_0)\mu(i_0) \quad (1.1.14)$$

From this we get

$$P(X_n = i / X_0 = j) = \sum_{i_{n-1}} p(i/i_{n-1})p(i_{n-1}/i_{n-2})\ldots p(i_1/i_0) = p^n(i/j) \quad (1.1.15)$$

$p^n(i/j)$ may be viewed as the $(i, j)^{th}$ entry of the matrix $P^n$, where $P = [p(i/j)]$. $P$ is a stochastic matrix, i.e. $p(i/j)$ satisfies

$$p(i/j) \geq 0, \sum_{i \in S} p(i/j) = 1.$$
Given the initial distribution of the Markov Chain $p_0$, the distribution of $X_n$ is then given by $\mu_n = P^n \mu_0$.

A Markov chain has a representation in terms of an oriented and weighted graph, which facilitates many discussions about its properties. The nodes of the graph are the states of the chain, and an arrow with weight $p(i, j)$ connects state $j$ to state $i$ if $p(i, j) > 0$.

1.1.3 Ergodic Chain and Invariant Distribution

Of special interest are the long-time properties of a Markov chain. In particular, we may ask: (i) Is there a unique invariant distribution? (ii) What are the convergence properties of an initial distribution $\mu_0$ towards this invariant distribution? When the state space $S$ is finite, these can be phrased as standard linear algebra questions. Indeed the existence and uniqueness of an equilibrium distribution is equivalent to the existence of a nonnegative eigenvector of $P$ with eigenvalue equal to 1 and multiplicity 1. Note that 1 is an eigenvalue of $P$ because of the condition $\sum_{i \in S} p(i/j) = 1$ for all $j$, which can be written, as $(1, 1, \ldots, 1)P = (1, 1, \ldots, 1)$.

Recall that if there exists a permutation matrix $Q$ such that

$$Q P Q^T = \begin{bmatrix} A_1 & B \\ 0 & A_2 \end{bmatrix}$$

(1.1.16)
then P is called reducible. Otherwise P is called irreducible. Perron-Frobenius theorem states that a irreducible positive matrix with spectral radius 1 (like P) has an eigenvalue 1 with multiplicity 1 and a strictly positive eigenvector. That is if P is irreducible, then, there exists a unique invariant distribution $\pi$ with strictly positive.

For any distribution $\mu_0$,

$$\mu_n = \tilde{P}_n \mu_0 \rightarrow \pi \quad (1.1.17)$$

where

$$\tilde{P}_n = \frac{1}{n} \sum_{j=1}^{n} P^j \quad (1.1.18)$$

A Markov chain whose stochastic matrix is irreducible is said to be ergodic. It implies that every state of the chain has a positive probability to be visited from every other state. In terms of the graph of the chain, it means that from any node one can follow arrows to reach any other node. A Markov chain can fail to be ergodic for two reasons (i) The chain may be disconnected; (ii) Some state of the chain may lead to others but may not be reachable from these. Generally, a non-ergodic Markov chain can be split into sub-chains, which, after appropriate removal of non-communicating states. This also means that one need to carefully distinguish existence of a unique invariant distribution and ergodicity of the Markov chain.
A stochastic process \( \{X_0, X_1, \ldots, X_n, \ldots\} \) at consecutive points of observation 0,1,...,n is a discrete time Markov chain (DTMC) if, for all \( n \in \mathbb{N}_0, X_n \in S \)

\[
P(X_{n+1} = x_{n+1} / X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} / X_n = x_n)
\]

(1.1.19)

Let \( S = \{0, 1, 2, \ldots\} \).

\[
p_{ij} = P(X_{n+1} = j / X_n = i) = P(X_1 = j / X_0 = i)
\]

(1.1.20)

are the one-step transition probabilities of a homogeneous chain, i.e., whose conditional probability mass function (pmf) is independent of time.

### 1.2 MONTE CARLO

Monte Carlo method is a simulation technique in which statistical distribution function is created by using a series of random numbers. This approach has the ability to develop many months, or years of data in a matter of few on a digital computer. The method is generally used to solve problems which cannot be adequately represented by mathematical models or where solution of the model is not possible by analytical method. Monte Carlo simulation yields a solution, which should be very close to the optimal, but not necessarily the exact solution. However it should be noted that this technique yields a solution that converges to the optimal or correct solution as a number of simulated trials lead to infinity.
1.2.1 Impact of Monte Carlo Methods on Scientific Research

Monte Carlo methods have been used for centuries, but only in the past few decades the technique has gained the status of a full-fledged numerical method capable of addressing the most complex applications. The Monte Carlo method may be thought of as similar to a political poll, where a carefully selected statistical sample is used to predict the behavior or characteristics of a large group.

Enrico Fermi in the 1930's used Monte Carlo in the calculation of neutron diffusion, and later designed the Fermiac, a Monte Carlo mechanical device used in the calculation of criticality in nuclear reactors. In the 1940's, Von Neumann, who established the mathematical basis for probability density functions (pdf), inverse cumulative distribution functions and pseudorandom number generators, developed a formal foundation for the Monte Carlo method. The work was done in collaboration with Stanislaw Ulam, who realized the importance of the digital computer in the implementation of the approach. Before digital computers were available to the labs, computer was a job title. Rows and columns of mathematicians did parallel computing. The applications, which arose mostly from the Manhattan Project, included design of shielding for reactors.
Uses of Monte Carlo methods have been many and varied since that time. In the late 1950's and 1960's, the method was tested in a variety of engineering fields. At that time, even simple problems were compute-bound. Many complex problems remained intractable through the seventies. With the advent of high-speed supercomputers, the field has received increased attention, particularly with parallel algorithms, which have much higher execution rates. All the pieces of the puzzle have just come into confluence for large-scale Monte Carlo analysis. First, supercomputers are now sufficiently powerful to enable the simulation of very large engineering and physics systems, involving thousands of surfaces and billions of particle emissions. Secondly, a comprehensive formulation for material properties exists in the aggregate of several models. Thirdly, an estimate of the number of trials required to achieve a specified level of accuracy is now obtainable prior to execution. This makes possible a formulation, which allows the number of emissions to evolve dynamically as the simulation proceeds. Finally, a number of investigators have effectively vectorized diverse Monte Carlo transport algorithms, with a sufficient base to establish a synthesized approach. We now even have a quantitative model, which allows the assessment of the degree of parallelism, and the amount of overhead required. Moreover, with the emergence of lagged Fibonacci generators, parallelization at any granularity appears to be easily implemented.
1.3 MARKOV CHAIN MONTE CARLO

Over the last few years MCMC has received considerable attention. MCMC is a computer intensive statistical tool, which has attracted the attention of many researchers. Using this theory it is easy to write very efficient algorithms for sampling from complicated target distributions. Hence it is easy to understand why these techniques have been applied in a vast number of different areas. Furthermore it has been around for a long time dating at least Metropolis et al. (1953) MCMC theory is in place. A good review of these aspects is given by Tierney (1994). The literature on MCMC methods is growing rapidly. The early 1990's have witnessed a burst of activities in applying MCMC in Bayesian methods to simulate Bayesian distributions. The simulation algorithm in its basic form is quite simple and becoming standards in many Bayesian applications. A good review is given by Gilks et al (1996). MCMC methods have been successfully used to overcome problems caused by missing data when using small networks for conventional statistics problems. All MCMC methods are ways to produce a stochastic process, which has a desired distribution as its stationary distribution. From the theory of stochastic process we infer that the empirical average of a function of the stochastic process will converge to the expectation of that function under the desired distribution.
MCMC is the idea of using simulations $X_1, X_2, \ldots, X_n$ of the Markov chain to approximate expectations

$$\mu = \mathbb{E}_\pi \{ g(X) \}$$  \hspace{1cm} (1.3.1)

by sample averages

$$\mu_n = \frac{1}{n} \sum g(X_i)$$  \hspace{1cm} (1.3.2)

where $\pi$ is equilibrium distribution also called invariant distribution, stationary distribution or ergodic limit of the Markov chain. In other words MCMC is just like Good Old-Fashioned Monte Carlo (GOFMC) except replace i.i.d in GOFMC by Markov chain to get MCMC and proceed mutatis mutandis.

It is highly nonobvious, even in its original incarnation, which was calculating ergodic limits for models of physical systems (Metropolis et al, 1953). What is obvious is run the (model of the) physical systems and average over time. Metropolis et al (1953) realized this and provided a simple algorithm for constructing a Markov chain having a specified equilibrium distribution. The Metropolis algorithm as generalized by Hastings (1970) and Green (1995) called the Metropolis–Hastings–Green (MHG) algorithm is the only known general method of MCMC. Every MCMC like method is a special case of the MHG algorithm.
1.4 HISTORICAL REVIEW

There are a number of isolated and undeveloped instances on much earlier occasions. In the second half of the nineteenth century a number of people performed experiments, in which they threw a needle in a haphazard manner onto a board ruled with parallel straight lines and inferred the value of \( \pi = 3.14 \) from observations of the number of intersections between needle and lines. In 1899, Raleigh showed that a one-dimensional random walk without absorbing barriers could provide an approximate solution to a parabolic differential equation. In 1936, Kolmogorov showed the relationship between Markov stochastic processes and certain integro-differential equations. In early part of the twentieth century, British statistical schools indulged in a fair amount of unsophisticated Monte Carlo work. Most of this seems to have been of didactic character and rarely used for research or discovery. Only on a few rare occasions was the emphasis on original discovery rather than comforting verification. In 1908 Student (Gosset) used experimental sampling to help him towards his discovery of the distribution of the correlation coefficient. In the same year Student also used sampling to bolster his faith in his so called t-distribution, which he had derived by a somewhat shaky and incomplete theoretical analysis. The real use of Monte Carlo methods as a research tool stems from work on the atomic bomb during the Second World War. This work involved a
direct simulation of the probabilistic problems concerned with random neutron diffusion in fissile material; but even at an early stage of these investigations, von Neumann and Ulam refined this particular Russian roulette and splitting methods. However, the systematic development of these ideas had to await the work of Harris and Herman Kahn in late 1940's. Fermi et al obtained Monte Carlo estimates for the eigen values of Schrodinger equation. During 1970's, the newly developing theory of computational complexity began to provide a more precise and persuasive rationale for employing the Monte Carlo method. The theory identified a class of problems for which the time to evaluate the exact solution to a problem within the class grows, at least, exponentially. The question to be resolved was whether or not the Monte Carlo method could estimate the solution to a problem in this intractable class to within a specified statistical accuracy in time upper bounded by a polynomial. Numerous examples now support this contention. Broder (1986) and Jerrum and Sinclair (1988) established the property for estimating, the number of perfect matching in a bipartite graph.

Numerical methods that are known as Monte Carlo methods can be loosely described as statistical simulation methods, where statistical simulation is defined in quite general terms to be any method that utilizes sequences of random numbers to perform the simulation. Monte Carlo is now used routinely in many diverse fields,
from the simulation of complex physical phenomena such as radiation transport in the earth's atmosphere and the simulation of the esoteric sub nuclear processes in high-energy physics experiments, to the mundane, such as the simulation of a Bingo game. The analogy of Monte Carlo methods to games of chance is a good one, but the game is a physical system, and the outcome of the game is not a pot of money or stack of chips (unless simulated) but rather a solution to some problem. The winner is the scientist, who judges the value of his results on their intrinsic worth, rather than the extrinsic worth of his holdings.

Statistical simulation methods may be contrasted to conventional numerical discrimination methods, which typically are applied to ordinary or partial differential equations that describe some underlying physical or mathematical system. In many applications of Monte Carlo, the physical process is simulated directly, and there is no need to even write down the differential equations that describe the behavior of the system. The only requirement is that the physical (or mathematical) system be described by a probability density function (pdf). Once the pdf's are known, the Monte Carlo simulation can proceed by random sampling from the pdf's. Many simulations are then performed and the desired results are taken as an average over the number of observations. In many practical applications, one can predict the statistical error in this average result, and hence an
estimate of the number of Monte Carlo trials that are needed to estimate the given error.

1.5 MCMC ALGORITHMS

The two most commonly used algorithms in MCMC applications are (i) Metropolis Algorithms and (ii) Gibbs Sampler. Geman and Geman (1984) presented the Gibbs sampler in context of spatial processes involving large number of variables e.g., Image reconstruction. They considered under which situations the conditional distributions given neighborhood subsets of the variables, uniquely determines the joint distribution. Besag (1974) has shown that if the joint distribution \( P(\theta_1, \theta_2, ..., \theta_d) \) is positive over its entire domain, then the joint distribution is uniquely determined by the \( d \) conditional distributions

\[
P(\theta_1 / \theta_2, ..., \theta_d), ..., P(\theta_d / \theta_1, ..., \theta_{d-1}).
\]

Li (1988) applied the Gibbs sampler in the context of multiple imputation which is elaborately discussed with application in Chapter V in the context of MCMC. Li suggested that the complete data be partitioned in to \( d+1 \) parts, \( X_0, X_1, ..., X_d \), where the observed data is \( X_0 \) and \( X_1, ..., X_d \) is a partition of missing data. Li assumes that \( X_i \) can be sampled from \( P( X_i / X_j, j \neq i ) \). His algorithm is as follows:
Step 1: Sample \( X_1^{(0)}, \ldots, X_d^{(0)} \) from some distributions.

Step 2: Sample \( X_1^{(i)} \) from \( P(X_1/ X_0, X_2^{(i-1)}, \ldots, X_d^{(i-1)}) \)

Sample \( X_2^{(i)} \) from \( P(X_2/ X_0, X_1^{(i)}, X_3^{(i-1)}, \ldots, X_d^{(i-1)}) \)

Sample \( X_d^{(i)} \) from \( P(X_d/ X_0, X_2^{(i)}, X_3^{(i)}, \ldots, X_{d-1}^{(i)}) \)

Step 2 is then cycled until the algorithm converges. Li suggested that multiple paths be considered to check for convergence. He also illustrated the methods in the context of categorical data, latent variables, and censored life data and provided conditions such that the distribution of \( (X_1^{(i)}, \ldots, X_d^{(i)}) \) converges to \( P(X_1, X_2, \ldots, X_d) \) geometrically faster. Like Metropolis et al. (1953) and Geman and Geman (1984), Li (1988) represented the process as a Markov chain with the joint posterior distribution as the stationary distribution.

Tanner and Wong (1987) presented the data augmentation algorithm, which is a two-component version of the Gibbs sampler. This is further discussed elaborately in Chapter II. One of the basic contributions of Tanner and Wong (1987) was to develop the framework in which the Bayesian can be performed in the context of iterative Monte Carlo algorithms. Moreover, in their rejoinder they sketched a Gibbs sampler approach for handling hierarchical models with t errors. Gelfand and Smith (1990) presented a review of data augmentation, the Gibbs sampler and the Sampling Important
Resampling (SIR) algorithm due to Rubin (1987). These authors applied the approaches to several examples and provided an initial practical comparison of the methods. Many researchers discussed the applications of the Gibbs sampler to a variety of important statistical problems (Gelfand et al (1990), Gelfand and Smith (1991), Carlin et al (1992), Carlin and Polson (1991), Gelfand et al (1992) and Carlin et al (1992)).

Hastings (1970) suggested a generalization of the Metropolis algorithm. He illustrated how to use this algorithm to simulate Poisson and Normal deviates as well as random orthogonal matrices. In this section we show how one can use the Metropolis algorithm to construct a Markov chain with equilibrium distribution $\pi(x)$.

We first consider the idea in the discrete case. Let $Q = \{ q_{ij} \}$ be a specified symmetric transition matrix. At a given step, randomly draw state $s_j$ from the $i$th of row of $Q$. With known probability $a_{ij}$, move from $s_i$ to $s_j$, otherwise, remain at step $s_i$. This construction defines a Markov chain with transition matrix $p_{ij} = a_{ij}q_{ij}$ ($i \neq j$) and

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}.$$ 

Following Metropolis et al (1953), we let

$$a_{ij} = \begin{cases} 1, & \text{if } \frac{\pi_j}{\pi_i} \geq 1 \\ \frac{\pi_j}{\pi_i}, & \text{if } \frac{\pi_j}{\pi_i} < 1 \end{cases} \quad (1.5.1)$$
This chain is reversible, since

\[
\pi_i p_{ij} = \pi_i \min \left\{ 1, \frac{\pi_j}{\pi_i} q_{ij} \right\} = \min \{ \pi_i, \pi_j \} q_{ij} = \min \{ \pi_i, \pi_j \} q_{ji} = \pi_j p_{ji} \quad (1.5.2)
\]

The equilibrium distribution will be unique if \( Q \) is irreducible (Ripley 1987). A sufficient condition for convergence (if \( \pi \) is not constant) is being able to move from any state to any other under \( Q \) (Ripley 1987). Following Barker (1965)

\[
\alpha_{ij} = \frac{\pi_j}{(\pi_i + \pi_j)} \quad (1.5.3)
\]

The resulting chain is reversible, since,

\[
\pi_i p_{ij} = \frac{\pi_i \pi_j}{(\pi_i + \pi_j) q_{ij}} = \pi_j p_{ji} \quad (1.5.4)
\]

Now we consider the idea in the continuous case: Here \( \pi \) is a density with respect to a measure \( \mu \) and \( f(x,y) \) is a symmetric transition probability function (i.e., \( f(x,y) = f(y,x) \)), then the Metropolis algorithm is given by

a) If the chain is currently at \( X_n = x \), then generate a candidate value \( y^* \) for next location \( X_{n+1} \) from \( f(x,y) \).
b) With probability

\[ \alpha(x, y^*) = \min \{ \pi(y^*) / \pi(x), 1 \} \]  

(1.5.5)

accept the candidate value and move the chain to \( X_{n+1} = y^* \). Otherwise reject and let \( X_{n+1} = x \). Thus the Metropolis algorithm yields a series of dependent realizations forming a Markov chain with \( \pi \) as its equilibrium distribution. A key observation is that the Metropolis algorithm only requires that \( \pi \) be defined up to the normalizing constant, since the constant drops out in the ratio \( \pi(y^*) / \pi(x) \). Tierney (1991) presents a number of suggestions for \( f(x, y) \). If \( f(x, y) = f(y-x) \), then the chain is driven by a random walk process. Possible candidates for \( f \) are the multivariate normal, multivariate t or split t (with a small of degrees of freedom). In situations where the multivariate t is used to generate candidate values, one would center the normal or the t at the current state of the chain \( x \), with the variance-covariance matrix possibly equal to some multiple of the inverse information at the posterior mode. Muller (1993) discussed the choice of scale issue in detail. Besides presenting the range of hybrid strategies by cycling / mixing different chains, Tierney (1991) presented formal conditions for convergence, rates of convergence, and limiting behavior of averages. Gelfand (1992) presented an analogue to the Gibbs stopper for the Metropolis algorithm.
A generalization of the Metropolis algorithm due to Hastings (1970) takes

\[
\alpha(x, y) = \min \left\{ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right\} \quad \text{if} \quad \pi(x)q(x, y) > 0
\]
\[
= 1 \quad \text{if} \quad \pi(x)q(x, y) = 0
\]

where \( q(x, y) \) is an arbitrary transition probability function. Note that if \( q \) is symmetric i.e., \( q(x, y) = q(y, x) \), as would be the case in using a multivariate normal or multivariate \( t \) to drive the algorithm, then the Hastings algorithm reduces to the Metropolis algorithm. Hastings (1970) considered the case where \( q(x, y) = q(y) \), which is closely related to importance sampling. Tierney (1991) called these as independence chains.

1.6 MOTIVATIONS AND SCOPE

We consider together some of the key issues in the application of Markov chain Monte Carlo methods. It is evident from a rapidly expanding literature that MCMC is applicable to a very wide range of complex models including Bayesian inference, many of which are at present well beyond the reach of other computational methods. Quite apart from philosophical differences, such formulations often have no obvious frequentist counterparts or there are no existing frequentist computational procedures although MCMC maximum likelihood might prove useful. The Bayesian paradigm especially persuasive in spatial applications where there are known to be local contextual regularities.
in the true scene that cannot be modeled plausibly by a physical process but for which one can (crudely) represent one's beliefs through a Markov random field. Thus, the true scene considered be fixing, rather than sampling from the process, but our views about it are represented stochastically.

The equivalence between any Markov random field formulation and the Gibbs sampler explains why the origins of such methods are to be found in spatial applications. One could describe Bayesian computation via MCMC abstractly as the restoration of hidden Markov random fields. Markov chain Monte Carlo has its uses even for models in which standard analytical or numerical methods are applicable. Thus, it can provide a (computationally intensive) check on the accuracy of other methods; its introduction may be necessary in carrying out sensitivity analysis; and it can be used to calculate complicated functionals of the posterior distribution, to which other methods may not relate. Indeed, one of the most appealing features of MCMC is the ease with which estimates of probabilities and associated quantities, such as credible intervals, are obtained directly from the corresponding empirical distributions, rather than via moment-based approximations. In this sense, we see MCMC as "putting the probability back into statistics" and have quoted posterior means and standard deviations because it is usual to do so, rather than out of conviction. Furthermore, functionals of the posterior
distribution whose estimators are sensitive to small changes in the MCMC sample need to be handled with great care; this can include moments as well as probabilities of rare events. Arguably, the most important aspect of MCMC in Bayesian inference is its flexibility, as referred to by Smith (1992); for many of us one of the most existing consequences of the combination of Markov Chain Monte Carlo methodology and ever-increasing computer power should be a model liberation movement Accordingly, one can invoke models that are considered most appropriate to the data, often involving nonstandard likelihoods and nonconjugate priors. MCMC methodology also deals rigorously with missing values, in properly representing the additional variability involved in the model. Whereas sensitivity to departures from the basic formulation may be with a suitably modified prior and/or likelihood, an alternative is to model uncertainty explicitly by adding appropriate additional layers to the existing hierarchy. Of course, the corresponding MCMC algorithm must be designed so as to move freely around those parts of the model space that have substantial support in the posterior distribution. Methodological development and practical experimentation will provide useful guidelines, as will probabilistic research into Markov chain simulation, the eventual choice will often depend on hardware and software environment and whether the context is routine production on many similar data sets or is a one-off analysis. The subsequent
chapters discuss formulation and use of MCMC methods and demonstrate the implementation, by analysis of experimental biomedical data in the presence of heterogeneity. The applications discussed in the subsequent chapters will also compare MCMC based estimates derived from posterior distributions with maximum likelihood estimates. However the maximum likelihood estimates can be difficult to obtain involving much integration. In such cases the MCMC methods offer substantial computational advantages.

The main aim and objective of this work are

1. To collect and unify MCMC methods with special reference to Bayesian Inference.

2. To establish some new MCMC Methodologies for non normal, non linear models.

3. To compare the performance of different MCMC algorithms.

4. To provide a Coherent background to the direction in which MCMC biomedical applications are moving.