

# MARKOV CHAIN MONTE CARLO

---

## STOCHASTIC SIMULATION FOR BAYESIAN INFERENCE

---

**Dani Gamerman**

*Professor of Statistics  
Federal University of Rio de Janeiro  
Brazil*

---



**CHAPMAN & HALL/CRC**

---

A CRC Press Company

Boca Raton London New York Washington, D.C.

# Introduction

---

## Overview of Bayesian Inference

This is a book about Statistical Inference, the area of Science devoted to drawing conclusions or inference about data through quantitative measurements. There is often uncertainty associated with measurements, either because they are made with imprecise devices or because the process under which these quantifications become available is not entirely controlled or understood. The tool used to quantify uncertainties is probability theory and probability distributions are associated with uncertain measurements. The specification of probability distributions to the uncertain measurements or random variables in a given problem along with possible deterministic relations between some of them defines a statistical model.

An example considered later in the book (see Example 6.4) is the study of the impact advertising expenditure may have on sales of a product or some surrogate measurement such as advertising awareness. In this case, the uncertain measurements are the results from weekly opinion polls carried out in a given population of interest. The result of an opinion poll is given by the percentage  $y$  of people who remembered having watched the advertisement on TV. It is expected that advertising expenditure  $x$  might have an effect on awareness. Therefore a deterministic relation is established to link its effect on the awareness probability  $\pi$ . The simplest link is given by the linear relation  $\pi = \alpha + \beta x^*$ . Since  $\pi \in [0, 1]$ , it is usual in such cases to transform it to the real line before equating it to the linear form. A very common transformation is

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$$

In either case, the larger the value of  $\beta$ , the more effective the advertisement campaign is in boosting awareness. There are, of course, many possible relations that can be entertained. The collection of the possible relations along with probability specifications for the percentages from the polls defines a statistical model.

Once a model is built, there are many ways to proceed with inference. The Bayesian approach considers uncertainties associated with all unknown

\* The value of the expenditure should take into account the instantaneous expenses but also the downweighted values of the expenses over previous weeks.

quantities whether they are observed or unobserved. Inference is drawn by constructing the joint probability distribution of all unobserved quantities based on *all* that is known about them. This knowledge incorporates previous information about the phenomena under study and is also based on values of observed quantities, when they are available. This book assumes the general case where both pieces of information are available.

In this case, the distribution of *unknowns* given the *knowns* is called the *posterior distribution* because it is obtained *after* the data is observed. The unknown quantities may include future observations (that are currently unknown). Inference about them is referred to as prediction and their marginal distribution is referred to as the predictive distribution. The operations required to obtain these distributions are derived in Chapter 2 and exemplified in a number of typical situations that occur in practice.

For the example, the quantities of interest are  $\alpha$  and  $\beta$ , used to define a link between expenditure and awareness. They are unknown, otherwise there would be little point in performing the polls. Marketing experience may provide some background information on them. Another source of information is provided by the result of the polls once they are carried out and the percentages become known. Bayesian inference provides the tools to combine these pieces of information to obtain the posterior distribution of  $\alpha$  and  $\beta$  based on previous, background knowledge and the observed information from the polls. There may also be interest in predicting future results from the polls for an anticipated advertisement expenditure in future weeks. In this case, the future results are added to the set of unknown quantities of interest.

Obtaining the posterior distribution is an important step but not the final one. One must be able to extract meaningful information from this distribution and translate it in terms of its impact on the study. This is mainly concerned with evaluation of point summaries such as mean, median or mode, or interval summaries given by probability intervals. In a few examples, this extraction or summarization exercise can be performed analytically, which means that an exact appraisal of the situation can be made. These cases are also illustrated in Chapter 2 and for them, the inferential task is completed.

In the advertisement study, the main interest is the evaluation of the value of  $\beta$ . If its distribution is concentrated with large probability around positive values then the study confirms that advertisement boosts awareness. Quantification is also important: the larger the values of  $\beta$ , the better the advertisement campaign is in raising awareness about the product.

In most cases, however, the complexity of the model prevents this simple operation from taking place. The complexity is sometimes caused by the combination of the sources of information available for a given quantity. In other cases, it is caused by the sheer amount of quantities required for an adequate description of the phenomena studied. In some cases, it may

even be caused by a combination of many quantities with many sources of information for some of them.

In the example, a more adequate description of the process is provided by a model that allows the links between expenditure and awareness to change with time. Different habits, changing environment, other rival advertisement campaigns and change in advertisement campaign are all reasons for a dynamic modelling. One possible representation is to allow the quantities  $\alpha$  and  $\beta$  to vary as time passes. The relation between the awareness probability  $\pi_t$  and the expenditure  $x_t$  at week  $t$  becomes

$$\text{logit}(\pi_t) = \alpha_t + \beta_t x_t$$

where the unknown quantities  $\alpha_t$  and  $\beta_t$  are now allowed to change with the week. This automatically leads to a substantial increase in the number of the unknown quantities. Also, one must expect some degree of similarity between links in adjacent weeks. One convenient form to specify similarities is

$$\begin{aligned}\alpha_t &= \alpha_{t-1} + w_1 r_t \\ \beta_t &= \beta_{t-1} + w_2 r_t\end{aligned}$$

Note that the number of unknown quantities has risen dramatically from 2 to  $2n$  where  $n$  is the number of weeks considered in the study. The incorporation of these similarities in the model also means that the structure of the model has increased in complexity. The distribution of the unknown quantities has consequently become more complex to handle.

One is inevitably led to seek approximations that can provide at least a rough guide to the exact but unobtainable answer. There are many ways to tackle this problem and a variety of suggestions have been proposed in the literature, with more emphasis on this aspect from the 80s. The timing is related to increased computing power enabling more sophisticated and computationally-based solutions. These solutions can be broadly divided into two groups: deterministic and stochastic approximations.

Some of the deterministic methods are based on analytical approximations whereas others are based on numerical approximations. They have received a great deal of attention in the literature and have been applied with success in problems where the number of unknown quantities is small. Chapter 3 reviews the main approximating techniques, pointing at their strengths and weaknesses.

An entirely different perspective to extracting relevant information contained in a given distribution is provided by stochastic simulation. The approach here is to use values simulated from the distribution of interest. A collection of these values forms a sample and defines a discrete distribution concentrated on the sample values. The distribution of these values is an approximation to the parent distribution used for the simulation. Then, all relevant calculations with the parent distribution can be approxi-

matly made with the sample distribution. In particular, the sample can be grouped into intervals and the histogram of relative frequencies plotted. If a large number of these values is simulated then the resulting histogram will be a very close approximation to the density of the distribution of interest. Chapter 3 also describes approximating techniques based on stochastic simulation.

Stochastic simulation, or Monte Carlo, techniques have a few attractive features that may explain their recent success in Statistical Inference. First, they have strong support in probability results such as the law of large numbers (equation (3.8)). It ensures that the approximation becomes increasingly better as the number of simulated values increases. This number is controlled by the researcher and only time and cost considerations may prevent a virtually error-free approximation. Also, at any stage of the simulation process, the approximation error may be probabilistically measured using the central limit theorem (equation (3.7)).

The main thrust of the book is the description of techniques devoted to perform Bayesian inference based on stochastic simulation, hence its subtitle *Stochastic simulation for Bayesian inference*. Before applying simulation, it is important to present basic, direct simulation operations to those not familiar with them. This is the purpose of Chapter 1. Many of the results presented there will be returned to in a more elaborate setting in later chapters.

Using these techniques, it is possible to devise simulation schemes to draw values from the distribution of  $\alpha$  and  $\beta$  (in the static model setting) but they do not provide adequate solutions to the more elaborate case of time-varying  $\alpha_t$  and  $\beta_t$ . These techniques will tend to be very inefficient as the dimension of unknown quantities increase, and more sophisticated simulation techniques will have to be used.

### Overview of MCMC

Nowadays, there are many problems of interest that fall into the category of large dimension models. Dynamic settings are just an example. Other examples also arise in the context of hierarchical or random effects models and models for spatial data. The first group roughly deals with unstructured additional variation whereas the second group deals with variations due to a neighbouring structure. They will also be considered in later chapters. Models with measurement errors and a mixture or combination of models are also settings for large dimension models.

The title of the book, *Markov Chain Monte Carlo*, refers to an area of Statistics, usually referred to as MCMC by taking the first letters of each word. MCMC will be described in detail in this book. It provides an answer to the difficult problem of simulation from the highly dimensional distribution of the unknown quantities that appear in complex models.

In very broad terms, Markov chains are processes describing trajectories where successive quantities are described probabilistically according to the value of their immediate predecessors. In many cases, these processes tend to an equilibrium and the limiting quantities follow an invariant distribution. MCMC techniques enable simulation from a distribution by embedding it as a limiting distribution of a Markov chain and simulating from the chain until it approaches equilibrium. Before understanding simulation through Markov chains, or MCMC in short, it is important that properties of Markov chains are well understood. For the sake of those not familiar with them, Chapter 4 reviews the most relevant results.

The introduction of Markov chains in the simulation schemes is vital. It allows handling of complicated distributions such as those arising in the large dimension models mentioned above. It is interesting that introduction of an additional structure, the Markov chain, into an already complex problem ends up solving it! There is also the matter of the extra work involved in simulation of a single value by MCMC: a complete sequence of values of a chain until it reaches equilibrium is required and only the equilibrium value can be taken as a simulated value from the limiting distribution. Fortunately, there are also analogues of the law of large numbers and central limit theorems (equations (4.6) and (4.9), respectively) for Markov chains. They ensure that most simulated values from a chain can be used to provide information about the distribution of interest.

There is still the question of how to build a Markov chain whose limiting distribution is exactly the distribution of interest, namely the distribution of all the unknown quantities of the model. It is amazing that not only is this possible but that there are large classes of schemes that provide these answers. One such scheme is Gibbs sampling. It is based on a Markov chain whose dependence on the predecessor is governed by the conditional distributions that arise from the model. It so happens that many models have a complex joint distribution but by construction (some of) their conditional distributions are relatively simple. Gibbs sampling explores this point and is able to provide simple solutions to many problems. Gibbs sampling is presented in Chapter 5 and exemplified in a number of situations including models with hierarchical structure and models with a dynamic setting.

There are many ways that MCMC can be used in any given situation. The main concern is efficient computation. Efficiency can be measured by the ease with which a simulated sample is obtained. It takes many aspects into consideration, such as choice of conditional distributions to use, need for transformations of the quantities, cost and time of a simulation run and stability of the solutions obtained. These matters are also dealt with in Chapter 5.

Another scheme is given by the Metropolis-Hastings algorithms, presented in Chapter 6. They are based on a Markov chain whose dependence on the predecessor is split into two parts: a proposal and an acceptance of

the proposal. The proposals suggest an arbitrary next step in the trajectory of the chain and the acceptance makes sure the appropriate limiting direction is maintained by rejecting unwanted moves of the chain. They provide a solution when even the conditional distributions of interest are complex, although their use is not restricted to these cases. Metropolis-Hastings algorithms may come in a variety of forms and these can be characterized and studied. Some of their forms may be seen as generalizations of Gibbs sampling.

Going back to the example, the conditional distributions of  $\alpha_i$  and  $\beta_i$  have also proved to be complex and Metropolis-Hastings algorithms seem to be a natural choice. Many schemes can be contemplated and a few of them are selected for numerical comparison and presented in Chapter 6.

It should also be noted that Bayesian Inference is not necessarily completed after summarizing information about unknown quantities of a given model. There may be other relevant operations to perform such as model evaluation and model comparison, involving more than one model. Of particular interest is the joint consideration of a (large) number of possible models. Bayesian Inference and MCMC can be accommodated to handle these questions. Alternative models can also be used as auxiliary devices in designing a MCMC method for a particular model. All these points are covered in Chapter 7.

## Notation

Whenever possible, the same notation is maintained throughout the book. Distributions are identified with their density or probability functions and variables are generically treated as if they are continuous. Posterior densities are denoted by  $\pi$  and their approximations (described throughout the book) by  $q$ , observed quantities by roman letters  $x, y, \dots$  and unobserved quantities or parameters by greek letters  $\theta, \phi, \dots$ . No distinctions are made between a random variable and its observed value and between scalar, vector and matrix quantities although matrices are generally denoted by capital letters and scalar and vector quantities by lower case letters. Vectors are always arranged in a column unless otherwise stated. The transpose of a vector  $x$  is denoted by  $x'$  and its dimension generally denoted by  $d$ .

The complement of an event  $A$  is denoted by  $\bar{A}$ , the probability of an event  $A$  is denoted by  $Pr(A)$ , and expectation and variance of a random quantity  $x$  are respectively denoted by  $E(x)$  and  $Var(x)$ . The covariance and correlation between random quantities  $x$  and  $y$  are respectively denoted by  $Cov(x, y)$  and  $Corr(x, y)$ . The number of elements of a set  $A$  is denoted by  $\#A$ . The indicator function is denoted by

$$I(x \in A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

Approximations are denoted by a  $\cdot$  superimposed to the relevant symbol. Therefore,  $\hat{=}$  stands for approximately equal,  $\hat{\sim}$  stands for approximately distributed as and  $\hat{\propto}$  stands for approximately proportional to.

Components of a vector  $x$  of fixed dimension are denoted by  $x_1, x_2, \dots$  whereas elements of a sequence  $x$  will tend to be denoted by  $x^{(1)}, x^{(2)}, \dots$ . This will help to distinguish between the component and the sequence dimensions when dealing with vector sequences. The identity and diagonal  $d \times d$  matrices are respectively denoted by

$$I_d = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix} \quad \text{and} \quad \text{diag}(c_1, \dots, c_d) = \begin{pmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & c_d \end{pmatrix}$$

The absolute value of the determinant of a matrix  $A$  is denoted by  $|A|$ .