AN INTRODUCTION TO BAYESIAN STATISTICS AND MCMC

CHAPTER 4

MCMC

> Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell *(Biometrika,* Vol.I.p.219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation *f* of each sample determined.
>
> **William Searly Gosset** (“**Student**”), 1908.

4.1. Samples of marginal posterior distributions

4.1.1. *Taking samples of marginal posterior distributions*

We have seen in chapter 2 that two great advantages of Bayesian inference are *marginalisation* and the possibility of calculating actual *probability intervals*. Integrals should be performed for both marginalizing and obtaining these intervals. This does not represent a problem for very simple models, but the difficulty increases when models have several effects and different variance components. These difficulties stopped the progress of Bayesian inference for many years, and often the only practical solution was to find a multivariate mode, renouncing to the possibility of marginalisation. Even obtaining the precision of these modes was an impossible task in many circumstances, because it was also necessary to calculate integrals to find credibility intervals. Most of these problems disappeared when a system of integration based in random sampling of Markov chains was made available. Using these methods, we do not obtain the posterior marginal distributions, but just random samples from them. This may look disappointing, but has many advantages as we will see soon.

Let us put an example. We need to find the marginal posterior distribution of the difference between the treatments S and C for the meat quality trait “flavour intensity”, given the data, measured by a panel test in a scale from 1 to 5. We are going to estimate this distribution by obtaining random samples of the marginal posterior distribution of the treatments given the data. We obtain two lists of random numbers

$$f(S \mid \mathbf{y}): [3.1, 3.3, 4.1, 4.8, 4.9, \ldots]$$
$$f(C \mid \mathbf{y}): [2.4, 2.6, 2.6, 2.6, 2.8, \ldots]$$

since both lists are random samples of the marginal posterior distributions of the effects, the difference sample by sample (i.e.; 3.1–2.4 = 0.7, 3.3–2.6 = 0.7, 4.1–2.6 = 1.5, 4.8–2.6 = 2.2, etc.) is a list of numbers that are also a random sample of the difference between treatments.

f(S-C | **y**) : [0.7, 0.7, 1.5, 2.2, 2.1,…]

These lists are called *Markov chains*. SInce they are formed by random samples, they are called "*Monte Carlo*", just as the famous casino. We can make a histogram with these numbers and obtain an approximation of the posterior distribution of S–C given the data **y** (figure 4.1).

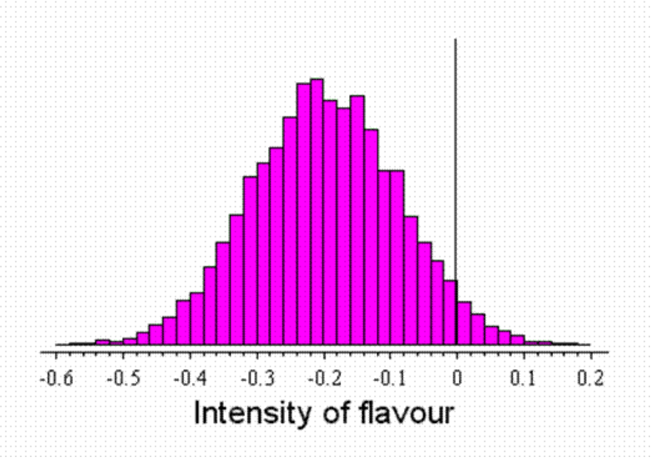

Figure 4.1 A histogram made by randomly sampling the posterior distribution of the difference between two treatments f(S–C | **y**) for 'Intensity of flavour'.

From this random sample it is easy to make Bayesian inferences as we will see later. For example, if we want to estimate the mean of this posterior distribution we just calculate the average of the *chain* of numbers sampled from f(S–C | **y**).

This chain of sampled numbers from the posterior distribution can be as large as we want, thus we can estimate the posterior distribution as accurately as we need. There is a sampling error that depends on the size of the sample, but also on how correlated the samples are. For example, if we take 500 samples and the correlation between them is 1 we do not have 500 samples because it is always the same one. The "effective number" that we have can be calculated; i.e., sample size of uncorrelated numbers that estimates the posterior distribution with the same accuracy as with our current chain.

Another important point is that we can directly sample form marginal distributions. If we find a way to obtain random samples (x_i, y_i) of a joint posterior distribution f(x,y), each x_i is a random sample of the marginal distribution f(x) and each y_i is a random sample of the marginal distribution f(y).

4.1.2. *Making inferences from samples of marginal posterior distributions*

From a chain of samples, we can make inferences. Let us take the former example, in which we have a chain of random samples of the posterior distribution for the difference between the selected and the control population. We now have a chain of 30 samples. Let us order the chain from the lowest to the highest values

f(S–C | **y**) : [-0.5, -0.4, -0.2, -0.1, -0.1, 0.0, 0.0, 0.1, 0.2, 0.2, 0.2, 0.2, 0.5, 0.6,
0.6, 0.7, 0.7, 1.1, 1.1, 1.3, 1.5, 1.8, 1.8, 1.8, 1.8, 2.0, 2.0, 2.1, 2.2, 2.4]

Now we want to make the following inferences:

1) *What is the probability of S being higher than C?* (Figure 2.8.a)

$$P(S>C) = P(S–C>0)$$

We estimate this probability counting how many samples higher than zero we have and divide by the total number of samples. We have 23 samples higher than zero from a total of 30 samples, thus our estimate is

$$P(S>C) = \frac{23}{30} = 0.77$$

2) *What is the probability of the difference between groups being higher than a relevant value that we determine to be 0.5?* (Figure 2.10.a)

We count how many samples are higher than 0.5. We find 17 samples higher than 0.5. Since we have 30 samples, the probability of the difference between groups being higher than 0.5 is

$$P(S-C>0.5) = \frac{17}{30} = 0.57$$

3) *What is the probability of the difference between the groups being different from zero?* (Figure 2.11.a)

Strictly speaking, the probability of being different from zero is 1, since this difference will never exactly take the value 0.000000……., thus we have to reformulate the question. We should define the minimum value of this difference from which lower values will be considered, in practice, as null. This is the "Relevant value" that we have defined in chapter 2. The question now should be:

What is the probability of similitude between the groups?

or

What is the probability of the difference between the groups being irrelevant (like zero for practical purposes)?

We decide, basing our decision in our knowledge of the problem, that a *relevant* difference will be any one equal or higher than ± 0.1. We see that only two samples are lower than 0.1 and higher than -0.1, thus

$$P(|S-C| \geq \text{Relevant value}) = \frac{2}{30} = 0.07$$

4) *What is the probability of the difference between groups being between 0.1 and 2.0?* (Figure 2.3)

We have 20 samples between both values (including them), thus this probability is

$$P(0.1 \leq S\text{–}C \leq 2.0) = \frac{20}{30} = 0.67$$

5) *What is the minimum value that the difference between treatments can take, with a probability of 70%*? (Figure 2.9.a)

We look for a *guaranteed value* with a probability of 70%, as defined in chapter 2. Let us take the *last* 70% of the samples of our ordered chain. 70% of 30 samples is 21 samples, thus we take the *last* 21 samples of the chain. The first value of this set, which is the lowest one as well, is 0.2, thus we say that the difference between groups is at least of 0.2 with a probability of 70%.

6) *What is the maximum value that the difference between groups can take with a probability of 0.90?*

We take the *first* 90% of the samples of our ordered chain. 90% of 30 samples are 27 samples, thus we take the first 27 samples, and the highest value of this set (the last sample) is 2.0. Thus we say that the difference between groups will be 2.0 as a maximum with a probability of 90%.

7) *Whath is the shortest interval containing a 90% of probability?* (Figure 2.7.a)

The shortest interval (i.e., the most precise one) is calculated by considering all possible intervals containing the same probability. Since 90% of 30 samples is 27 samples, such an interval will contain 27 samples. Let us consider all possible intervals with 27 samples. These intervals are [-0.5, 2.0], [-0.4, 2.1], [-0.2, 2.2]. The first interval has a length of 2.5, the second one 2.5 and the third one 2.4, thus the shortest interval containing 90% probability is [-0.2, 2.2].

8) *Give an estimate of the difference between groups* (Figure 2.5)

Although it is somewhat illogical to say that this difference has a value just to immediately say that we are not sure about this value (and give an interval), it is usual to give point estimates of the differences between treatments. We have seen that we can give the mean, median or mode of the posterior distribution. The mean is the average of the chain, and the median the value in the middle, the value between the sample 15 and 16.

Estimate of the mean and median of the posterior distribution P(S-C):

Mean = $\frac{1}{30}\sum$ (-0.2, -0.2, -0.1, -0.1, -0.1, 0.0, 0.0, 0.1, 0.2, 0.2, 0.2, 0.2, 0.5, 0.6, 0.6, 0.7, 0.7, 1.1, 1.1, 1.3, 1.5, 1.8, 1.8, 1.8, 1.8, 2.0, 2.0, 2.1, 2.1, 2.2) = 0.86

$$\text{Median} = \frac{0.6 + 0.7}{2} = 0.65$$

To estimate the mode, we need to draw the distribution, since we have a finite number samples (it can happen, for example, that by chance we have few samples of the most probable value).

In this example, mode and median differ, showing that the distribution is asymmetric. Which estimate should be given is a matter of opinion, we should just be aware of the advantages and disadvantages, expressed in 2.2.1. Statisticians often prefer medians as point estimates. Medians have also the advantage of being robust to outliers; for example, take a sample of food conversion rate in pigs in which there is an error in the last data,

[2.3, 2.3, 2.3, 2.4, 2.5, 2.5, 24]

The mean of this set is 5.5, but the median is 2.4, much closer to what should be the real value. Sometimes the chains can sample outliers (particularly when we are making combinations of chains, for example finding ratios when the denominator is not far from zero). Medians are robust to these outliers.

4.2. Gibbs sampling

4.2.1. *How it works*

Now the question is how to obtain these samples. We will start with a simple example: how to obtain random samples from a joint posterior distribution f(x,z) that are also sets of samples from the marginal posterior distributions f(x), f(z). In chapter 5 we will estimate the marginal posterior distributions of the mean and the variance of a Normal distribution, in chapter 6 and 7 we will estimate marginal posterior distributions of the parameters of a linear model and in chapter 8 we will see how to estimate marginal posterior distributions of a variety of models.

We will use MCMC techniques, and a common one is called “Gibbs sampling”. What we need for obtaining these samples is:

1. Univariate distributions of each unknown parameter *conditioned* to the other unknown parameters; i.e., we need f(x|z) and f(z|x).
2. An algorithm for extracting random samples from these conditional distributions.

The first step is easy, as we have seen in chapter 3. The second step is easy if the conditional distribution has a recognisable form (Normal, Gamma, Poisson, etc.) for which we have algorithms that permits us to extract random samples. For example, to extract random samples from a Normal distribution

a) Take a random sample x between 0 and 1 from a random number generator (all computers have this).
b) Calculate

$$y = \sqrt{-2\log x} \cdot \cos(2\pi x)$$

Then y is a random sample of a N(0,1).

When we do not have this algorithm because the conditional distribution is not a known function or we do not have algorithms allowing us to extract random samples from it, other MCMC techniques can be used, but they are much more laborious as we will see later.

Once we have several conditional distributions from which we can sample, the Gibbs sampling mechanism starts as follows (Figure 4.2):

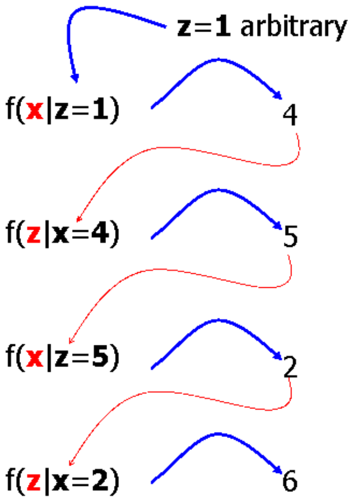

Figure 4.2. Gibbs sampling at work

1) Start with an arbitrary value for z, for example z=1
2) Extract one random sample from the conditional distribution f(x|z=1). Suppose this random sample is x=4
3) Extract one random sample from the conditional distribution f(z|x=4). Suppose this random sample is z=5
4) Extract one random sample from the conditional distribution f(x|z=5). Suppose this random sample is x=2
5) Extract one random sample from the conditional distribution f(z|x=4). Suppose this random sample is z=6
6) Continue with the process until obtaining two long chains

   x: 4, 2, …

   z: 5, 6, …

7) Disregard the first samples. We will see later how many samples should be disregarded and why.
8) Consider that the samples not disregarded are samples from the marginal distributions f(x) and f(z).

4.2.2. *Why it works*

Markov Chain Monte Carlo is a complex branch of the mathematics requiring a considerable effort for developing and understanding the main results. Exposing the mathematical methodology of this numeric methods is out of the scope of this book. Nevertheless, we will expose here intuitively the methods we need for our inferences, in order to understand how and why MCMC works. Consider the former example. Figure 4.3 shows $f(x,z)$ represented as lines of equal probability density (as level curves in a map).

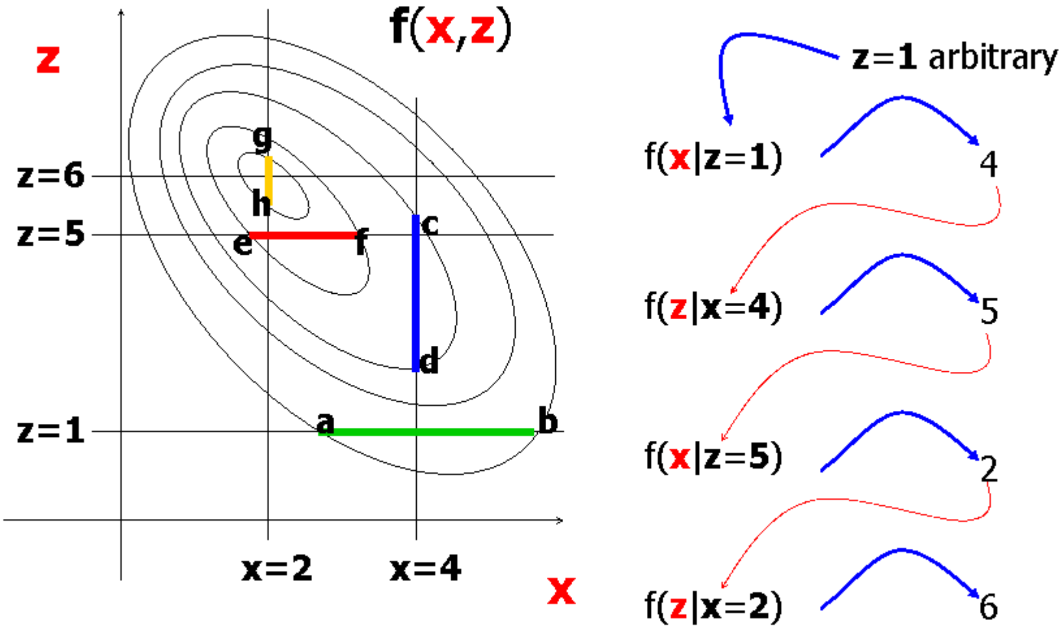

Figure 4.3. Gibbs sampling. The curves represent lines of equal probability.

Take an arbitrary value for z, say z=1. Sample a random number from the conditional density function f(x|z=1), which is the density function that has all possible values for x but all values of z are z=1. This function is represented in figure 4.1 as the line z=1, that has more probability density between ‘a’ and ‘b’ (in green) than in other parts of this line. Therefore, the number sampled from f(x|z=1) will be found between ‘a’ and ‘b’ more probably than in other parts of the conditional distribution. Suppose that the number sampled is x = 4.

Now sample a random number from the conditional density f(z|x=4). This distribution is represented in figure 4.1 as the line x=4, that has more density of probability between ‘c’ and ‘d’ (in blue) than in other parts of this line. Therefore, the number sampled from f(z|x=4) will be found between ‘c’ and ‘d’ more probably than in other parts of the conditional distribution. Suppose that the number sampled is z=5.

Now sample a random number from the conditional density f(x|z=5). This distribution is represented in figure 4.1 as the line z=5, that has more density of probability between ‘e’ and ‘f’ (in red) than in other parts of this line. Therefore, the number sampled from f(x|z=5) will be found between ‘e’ and ‘f’ more probably than in other parts of the conditional distribution. Suppose that the number sampled is x=2.

Now sample a random number from the conditional density f(z|x=2). This distribution is represented in figure 4.1 as the line x=2, that has more density of probability between ‘g’ and ‘h’ (in yellow) than in other parts of this line. Therefore, the number sampled from f(z|x=2) will be found between ‘g’ and ‘h’ more probably than in other parts of the conditional distribution. Suppose that the number sampled is z=6, we will carry on the same procedure until we obtain a chain of samples of the desired length.

Observe that we have the trend to sample from the highest areas of probability more often than from the lowest areas. At the beginning, z=0 and x=4 were points of the posterior distribution, but they were not random extractions, thus we were not interested on them. However, after many iterations, we will find more samples in the highest areas of probability than in the lowest areas, thus we will find random samples from the joint posterior distribution f(x,z) that are also random samples from the respective marginal posterior distributions f(x), f(z). This explains why the first points sampled should be discarded, and only after several cycles of iteration the samples are taken at random.

Notice that the conditional distributions f(x|z=1), f(x|z=5),… are different functions. For example, if x and z were the mean and variance of a Normal distribution, f(x|variance=1) is not the same distribution as f(x|variance=5). We will see a detailed example in chapter 5.

4.2.3. *When it works*

1) *Strictly speaking, it cannot be demonstrated that we are ultimately sampling from a posterior distribution.* A Markov chain must be *reducible* to converge to a posterior distribution. Although it can be demonstrated that some chains are not reducible, there is no general procedure to ensure reducibility.

2) Even in the case in which the chain is reducible, *it is not known when the sampling from the posterior distribution begins*. Even when having a reducible chain and the tests ensuring convergence, the distribution may not be stationary. Sometimes there are large sequences of sampling that give the impression of stability, and after many iterations the chains move to another area of stability.

The above problems are not trivial, and they occupy a part of the research in MCMC methods. Practically speaking, what people do is to launch several chains with different starting values and observe their behaviour. No pathologies are expected for a large set of problems (for example, when using multivariate distributions), but some more complicated models (for example, threshold models with environmental effects in which no positives are observed in some level of one of the effects) should be examined with care. By using several chains, we arrive to an iteration from which the variability among chains may be attributed to sampling error, and thus support the belief that samples are being drawn from the posterior distribution. There are some tests to check whether this is the situation (Gelman and Rubin, 1992). Another possibility is to use the same seed and different initial values; in this case both chains should converge and we can establish a minimum difference between chains to accept the convergence (Johnson 1996). When having only one chain, a common procedure is to compare the first part and the last part of a chain (Geweke, 1992). The biologist or the agricultural researcher living in a multivariate world should not have any problem with MCMC methods. For more complex models, good books for Bayesian inference with MCMC are Gelman et al. (2013) and Carlin and Louis (2008).

It should be noted that the special difficulties found when using MCMC methods for complex problems are similar in the classical statistical world. Finding a global maximum in multivariate likelihood with several fixed and random effects, out of the multivariate normal distribution world, is not an easy task usually. Complex models are difficult to handle under either paradigm. However, MCMC techniques transform multivariate problems in univariate approaches, and inferences are made using probabilities, having an easier interpretation.

4.2.4. *Gibbs sampling features*

To give an accurate description of the Gibbs sampling procedure used to estimate the marginal posterior distributions, in a scientific paper we should offer

In the Material and Methods section

1. *Number of chains:* When using several chains which converge, we have the psychological persuasion that no convergence problems were found. We have no limit for the number of chains; in simple problems, like linear models for treatment comparison, one chain is enough, but with more complex problems it is convenient to compute at least two chains. For

very complex problems, ten or more chains should be computed. For example, chains in figure 4.4.a indicates that in a complex problem convergence arrived, but figure 4.4.b questions convergence.

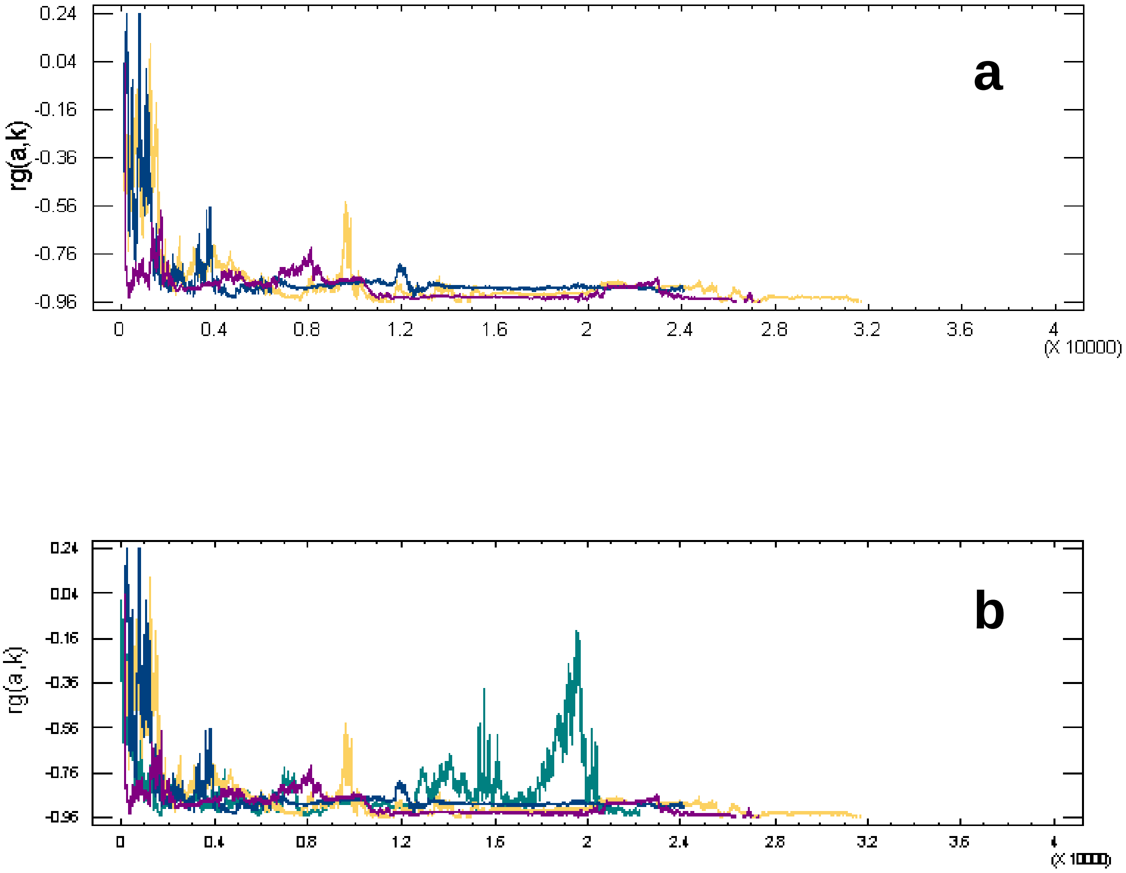

Figure 4.4 Several Gibbs sampling chains showing convergence (a) or not (b)

2. *Length of the chains*: It is customary to provide the length of the chains in order to have an idea about the complexity of the problem. Very long chains are performed when there are convergence problems or when all the samples are extremely correlated. For common problems like treatment comparisons, short chains of some thousands samples are long enough.

3. *Burn-in*: We have seen in 4.3.2 that the first samples are not taken at random and should be disregarded. When we start considering that the samples are random samples from the joint (and consequently from the

marginal) distributions is usually made by visual inspections of the chain. In many problems this is not difficult, and in simple problems like treatment comparison, convergence is raised after few iterations (figure 4.5). Although there are some methods to determine the burn-in (for example, Raftery and Lewis, 1992), they require the chain to have some properties that we do not know whether it actually has them, thus visual inspection is a common method to determine the burn-in.

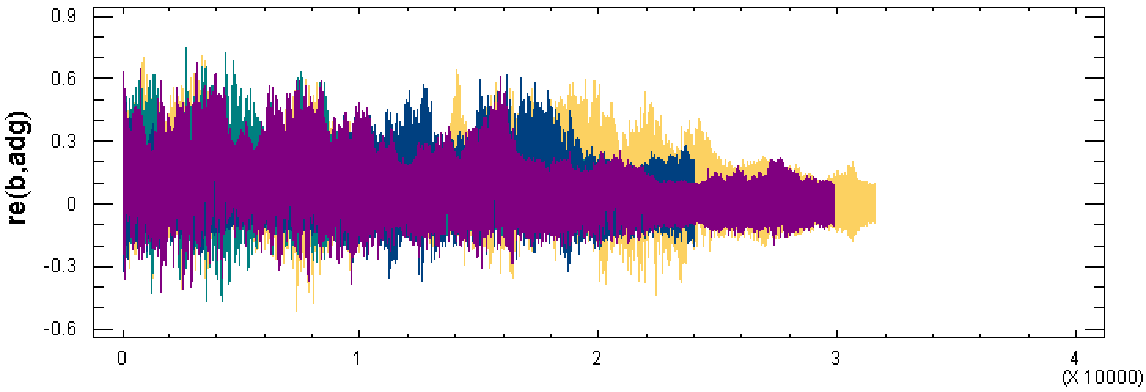

Figure 4.5 Several Gibbs sampling chains showing convergence from the first iterations

4. *Sampling lag*: We have seen in 4.2.2 how we started sampling in a part of the distribution, and how it is not probable to jump to the other side of the posterior distribution for a new sample; thus samples are often correlated. If the correlation between two successive samples is very high (say, 0.99) we will need more samples to obtain the same precision. For example, if the samples are independent, the sample mean has a variance which is the variance of the distribution divided by the number of samples (σ^2/n), but when the samples are correlated, this variance is higher because the covariances should also be taken into account. Collecting a high enough number of samples is not a problem, we can achieve the accuracy we desire, but in order to avoid collecting an extremely large number of highly correlated samples, consecutive samples are disregarded. For example, if only one of 100 samples is

collected, this substantially decreases the correlation between two collected consecutive samples. This is called the 'lag' between samples. Nevertheless, in common analysis in biology or agriculture, like treatment comparisons, no sapling lag is needed.

In Tables of results

5. *Convergence tests*: When having a positive answer from a convergence test we should say that "no lack of convergence was detected", because as we said before, there is no guaranty about the convergence. Some authors give the results of the tests. In common analyses, like treatment comparisons, this is not needed since convergence is obtained after a few iterations.

6. *Monte Carlo s.e.*): This is the error produced by the size of the sample. As we said before, it is not the same to estimate the posterior distribution with 500 samples than with 5,000 or 50,000, and the samples can also be correlated. This error is calculated using groups of samples and examining the sample means, or using temporal series techniques. Current software for MCMC gives the MCse. It can become as small as we want just by taking more samples of the posterior distribution. We should augment the sampling until this error becomes irrelevant (for example, when it is 10 times lower than the standard deviation of the posterior distribution).

7. *Actual sample size*: It is the equivalent number of independent samples that will have the same accuracy as the sample we have. For example, we can have 50.000 highly correlated samples that will lead to the same precision as 35 uncorrelated samples. The actual sample size gives us an idea about the real sample size we have, since having a high number of highly correlated samples does not provide much information. As we have said before, in many analyses, samples will not be highly correlated and actual sample size will not be commonly offered.

8. *Point estimates*: Median, mean, mode. When distributions are approximately symmetrical we should use one of them. Statisticians normally prefer the median for reasons explained in this book, but the mean and the mode are also used.

9. *Standard deviation of the posterior distribution*: When the distribution is approximately Normal, it is sufficient to know the s.d. for calculating HPDs; for example, HPD95% will be approximately twice the standard deviation.

10. *Credibility intervals:* HPD, guaranteed values [k, +∞) or (−∞, k]. We can give several intervals in the same paper; for example, guaranteed values for 80, 90 and 95% of probability.

11. *Probabilities*: P(S−C)>0, P(S−C>Relevant), Probability of similarity.

4.3. Other MCMC methods

Not always will we have an algorithm providing us random samples of the conditional distributions. In this case we can apply several methods. Here we can only outline some of the most commonly used. There is a whole area of research to find more efficient methods and to improve the efficiency of the existent ones. The readers interested in MCMC methods can consult Sorensen and Gianola (2002) or Gelman et al. (2013) for a more detailed account.

4.3.1. *Acceptance – rejection*

The acceptance-rejection method consists in covering the density f(x), from which we want to take random samples, with a known function g(x) (not a density distribution) having an algorithm allowing us to take random samples from it (Figure 4.6). By sampling many times and building a histogram, we can obtain a good representation of g(x). We call g(x) a “proposal function” because it is the one we propose for sampling.

Let us extract a random sample from g(x). Consider that, as in figure 4.6, the random sample x_0 extracted gives a value for $f(x_0)$ that is ½ of the value of $g(x_0)$. If we take many random samples of g(x) and build a histogram, half of them will be also random samples of f(x), but which ones? To decide this, we can throw a coin: “face” it is a random sample of f(x), “tail” it is not. If we do this, after sampling many times and building a histogram, we will have a good representation of f(x) at x_0.

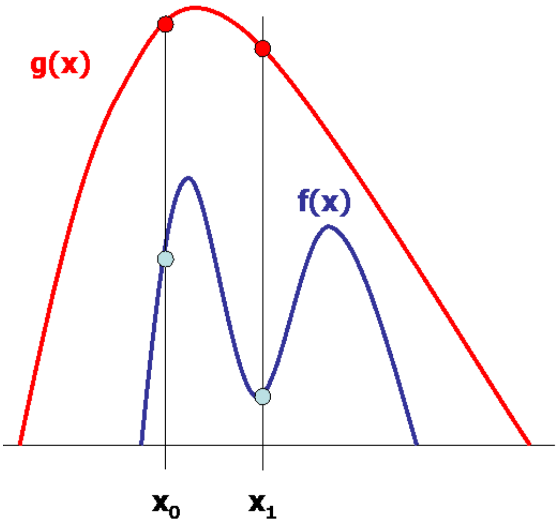

Figure 4.6. Acceptance – Rejection MCMC method. f(x) is the density from which we want to take random samples. g(x) is a function (not a density function) from which we know how to take random samples, covering f(x). x_0 is a random sample from g(x) that may also be a random sample from f(x)

We will proceed in a similar manner for the following sample, x_1, examining the ratio between $f(x_1)$ and $g(x_1)$. If this ratio is 1/6, as in figure 4.6, we can throw a die and decide that if we get a ‘1’ the sample is a random sample from f(x), otherwise it is not. Again, if we do this, when sampling many times and building a histogram, we will have a good representation of f(x) at x_1.

The general procedure is as follows:

1) Take a random sample x_0 from g(x)
2) Sample a number k from the uniform distribution U[0,1]
3) If $k < \frac{f(x_0)}{g(x_0)}$ accept x_0 as a random sample of f(x), otherwise, reject it.

For example, we take a random sample of g(x) and we get x_0 = 7. We take a random sample from U[0,1] and we get k=0.3. We evaluate the ratio of functions at x_0 and we obtain $\frac{f(7)}{g(7)} = 0.8 > 0.3$, thus we accept that 7 is a random sample of f(x).

How to find a good g(x) is not always easy. We need to be sure it covers f(x), thus we need to know the maximum of f(x). Some g(x) functions can be very inefficient and most samples can be rejected, which obliges to sample very many times. For example, in figure 4.7.a we have an inefficient function, x_0 is going to be rejected most of the times. We see in figure 4.7.b a better adapted function.

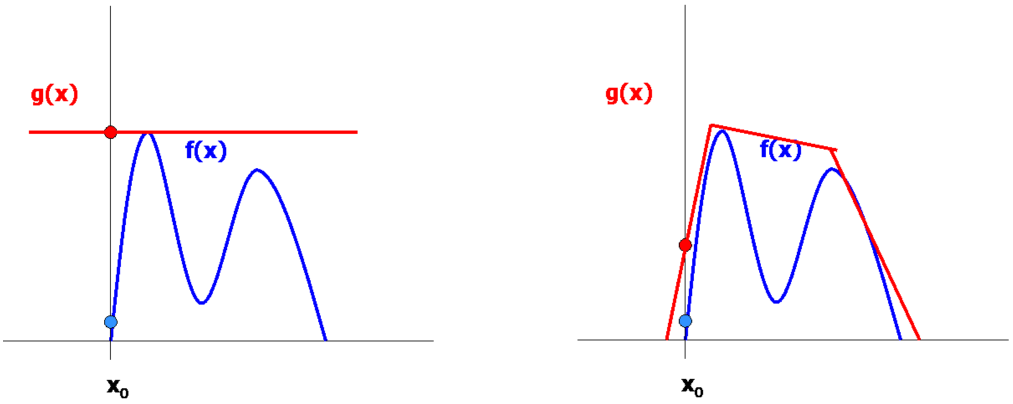

Figure 4.7. a. An easy but inefficient Accepting-Rejection function. **b.** A better adapted function

There are methods that search for a new g(x) according to the success we had in the previous sampling, they are called “adaptive acceptance rejection samplings”. As said before, there is a whole area of research in these topics.

4.3.2. *Metropolis-Hastings*

The method ([1]) has a rigorous proof based in the theory of Markov chains, that can be found for example in Sorensen and Gianola (2002). We will only expose here how it works.

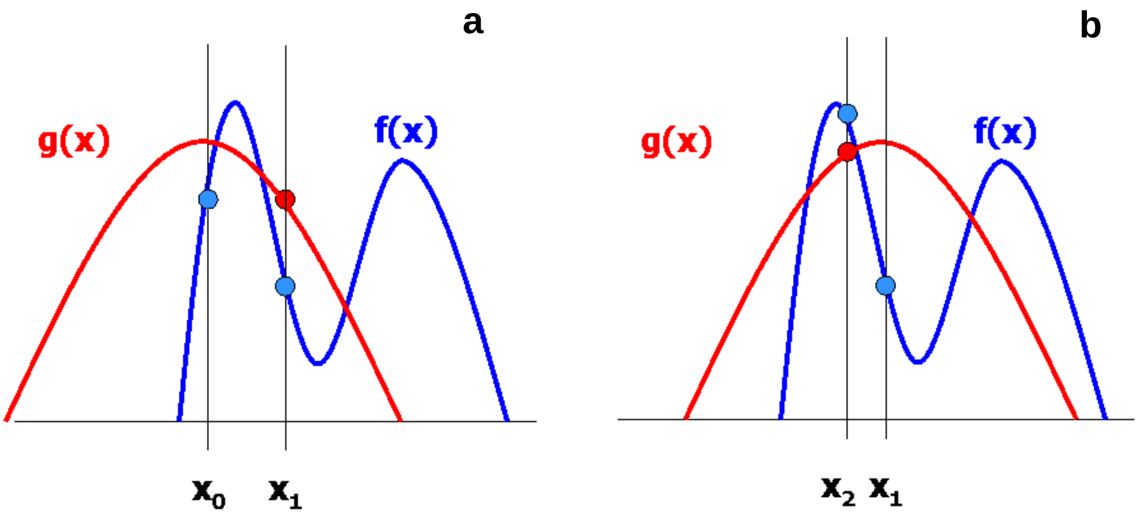

Figure 4.8. A proposal density g(x) from which we know how to take samples and the density function f(x) from which we need to take random samples. **a.** Start sampling. **b.** Moving the proposal density.

Our proposal function g(x) is now a density function ([2]), thus it should integrate to 1. We know how to take samples from g(x). The procedure works as follows:

1) Take a number x_0, arbitrary (figure 4.8a).
2) Sample x_1 from g(x)

[1] The following method was developed for symmetric proposal densities by Nicolas Metropolis in 1953, and it was improved for asymmetrical densities by Hastings (1970).

[2] Although we use the notation f(x) for density functions along this book, we will make an exception for g(x) in order to maintain the nomenclature used before for proposal functions.

3) If $f(x_1)>f(x_0)$ accept x_1 as a sample of f(x), otherwise, as in figure 4.6, use acceptance-rejection sampling. If x_1 is rejected, sample again from g(x) and repeat the process until you get a sample that is accepted.
4) If x_1 is accepted, *move* the mode of g(x) to x_1 as in figure 4.8.b, and sample again x_2 from g(x). Check whether $f(x_2)>f(x_1)$ and proceed as before.

Notice that by accepting x_1 when $f(x_1)>f(x_0)$, we tend to sample more often in the highest probability areas. We ensure we are sampling in the low probability areas by the acceptance-rejection method. When sampling enough times and constructing a histogram, it will reflect the shape of f(x).

A key issue of the method is to find the right proposal density g(x). It should be as similar as possible to the function f(x) from which we want to extract samples, in order to accept as many samples as possible when sampling. If we have a proposal density as in figure 4.8.a, we can accept many samples on the left part of f(x) but never move to the right part of f(x), and therefore we will not estimate f(x) but only a part of it. There is research on how to find efficient proposal densities, and there are adaptive Metropolis methods that try to change the proposal density along the sampling process to get more efficient g(x) densities.

Appendix 4.1

Software for MCMC

There is software that can be used in several mainframes and in Microsoft-Windows PCs allowing to analyse a large number of statistical models.

BUGS permits to make inferences from a large number of models. It is programmed in R-programming language and it allows adding R instructions. It is not charged by the moment, and it is widely used for Bayesian analyses. It can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs. It does not have the possibility of including the relationship matrix, but recently, Daamgard (2007) showed how to use BUGS for animal models.

TM is a Fortran90 software for multiple trait estimation of variance components, breeding values and fixed effects in threshold, linear and censored linear models in animal breeding. It has been developed by Luis Varona and Andrés Legarra, with the intervention of some other people. It is not charged and can be obtained from http://acteon.webs.upv.es/Software/TM.htm

Ignacy Mizstal has developed a set of programs under the name BLUPF90, covering a large amount of different problems, orientated to genetics. They can be downloaded from http://nce.ads.uga.edu/wiki/doku.php