CHAPTER 4

MCMC

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* (called credibility intervals by Bayesians). Both to marginalize and to obtain these intervals, integrals should be performed. For very simple models this is not a difficulty, but the difficulty increases when models have several effects and different variance components. These difficulties stopped the progress of Bayesian inference for many years. Often the only practical solution was to find a multivariate mode, renouncing to the possibility of marginalisation. But this mode was given without any error or measure of uncertainty, because it was also necessary to calculate integrals to find credibility intervals. Most of these problems disappeared when it was made available a system of integration based in random sampling of Markov chains. 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 selected and the control group for the meat quality trait "intensity of flavour", given the data, measured by a panel test in a scale from 1 to 5. But instead of this, we are going to obtain samples of the marginal posterior distribution of the selection and control effects given the data.

f(S | **y**): [3.1, 3.3, 4.1, 4.8, 4.9,…]
f(C | **y**): [2.4, 2.6, 2.6, 2.6, 2.8,...]

as both 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.) gives a list of numbers that is 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*. As they are formed by random samples, they are called "*Monte Carlo*". We can make a histogram with these numbers and obtain an approximation to the posterior distribution of S–C given the data **y** (figure 5.1).

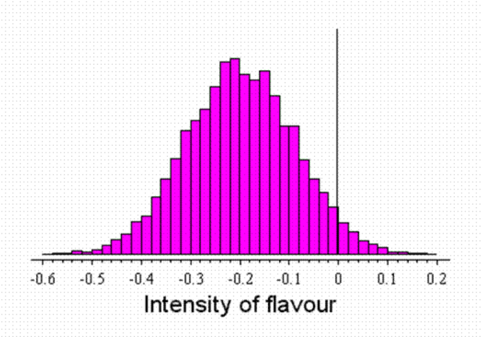

Figure 4.1 An histogram made by random sampling the posterior distribution of the difference between the selected and the control population f(S–C | **y**) for Intensity of flavour

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. Notice that it is not the same to estimate the posterior distribution with 500 samples than with 5,000 or 50,000. The samples can also be correlated. There is a sampling error that depends on the size of the sample but also on how correlated are the samples. For example, if we take 500 samples and the correlation between them is 1 we do not have 500 samples but always the same one. It can be calculated the "effective number" that we have; i.e., sample size of uncorrelated numbers that estimates the posterior distribution with the same accuracy that we do with our current chain of samples.

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). For example, if x and y can only take discrete values 1, 2, 3, … , 10, we take a chain of 500 samples of the joint posterior distribution and we order this sample according to the values of x, we have

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & & 2 & 2 & 2 & 2 & 2 & 2 & & 3 & 3 & & 10 \\ 1, & 1, & 1, & 2, & 2, & 3, & 4, & 4, & 4, & 5, & 5, & \ldots, & 1, & 1, & 2, & 3, & 3, & 3, & \ldots, & 1, & 1, & \ldots, & 10 \end{pmatrix}$$

We have seen that in the continuous case a marginal density f(x) is

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

The equivalent in the discrete case is

$$f(x_i) = \frac{1}{n_{ik}} \sum_k f(x_i, y_k)$$

where n_{ik} is the number of samples (x_i, y_k). We can see that to calculate the frequency of x=1 we take all pairs of values (1,1), (1,1), (1,2), (1,2), (1,3),… which is to take all possible values of (1,y_k) and sum. Thus the first row is composed by random samples of the marginal distribution f(x) and the second row is composed by random samples 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 have now a chain of 30 samples. Let us order the chain from the lowest to the highest values

f(S–C | **y**) : [-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]

Now we want to make the following inferences:

1) *Which is the probability of S being higher than C?* (Figure 2.6.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) *Which is the probability of the difference between groups being higher than 0.5?* (Figure 2.8.a)

We count how many samples are higher than 0.5. We find 17 samples higher than 0.5. As 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) *Which is the probability of the difference between groups being different from zero?* (Figure 2.9.a)

Strictly speaking, the probability of being different from zero is 1, since this difference will never take exactly the value 0.000000……., thus we have to define the minimum value of this difference from which lower values will be considered in practice as null. It is the same difference that is used in experimental designs when we decide that higher values will appear as significant and lower values as non significant. We call any higher value a *relevant* difference between groups or treatments. We decide 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

$$P(|S–C| \geq \text{relevant}) = \frac{2}{30} = 0.97$$

4) *Which is the probability of the difference between groups being between 0.1 and 2.0?* (Figure 3.3.b)

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

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

5) *Which is the minimum value that can take the difference between treatments, with a probability of 70%*? (Figure 2.7.a)

Let us take the *last* 70% of the samples of our ordered chain. A 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 0.2 with a probability of 70%.

6) *Which 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. A 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 as a maximum 2.0 with a probability of 90%.

7) *Which is the shortest interval containing a 90% of probability?* (Figure 2.5.a)

The shortest interval (i.e., the most precise one) is calculated by considering all possible intervals containing the same probability. As a 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 [-0.2, 2.0], [-0.2, 2.1], [-0.1, 2.2]. The first interval has a length of 2.2, the second one 2.3 and the third one 2.3, thus the shortest interval containing a 90% of probability is [-0.2, 2.0].

8) *Give an estimate of the difference between groups*

Although it is somewhat illogical to say that this difference has a value just to say later that we are not sure about this value and we should 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

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 and it is not possible to estimate with them accurately the mean (it ca happens, for example, that 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 just should know the advantages and disadvantages, expressed in 2.2.1.

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,y) that are also sets of samples from the marginal posterior distributions f(x), f(y). We will use MCMC techniques, and the most common one is called "Gibbs sampling" for reasons we commented in 1.1. What we need for obtaining these samples is:

1. To obtain univariate distributions of each unknown parameter *conditioned* to the other unknown parameters; i.e., to obtain f(x|y) and f(y|x).

2. To find a way to extract random samples from these conditional distributions.

The first step is easy, as we have seen in 3.4. The second step is easy if the conditional distribution have a recognisable form; i.e., they are Normal, Gamma, Poisson or other known distribution. We have algorithms that permit us to extract random samples of known distributions. For example,

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 is not a known function or we do not have algorithms to extract random samples from it, other MCMC techniques can be used, but they are much more laborious as we will se later.

Once we have conditional functions from which we can sample, the Gibbs sampling mechanism starts as follows (Figure 4.1):

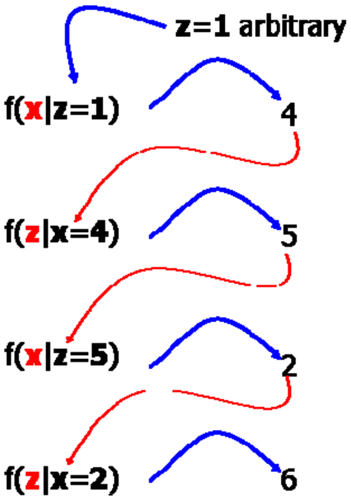

Figure 4.2. Gibbs sampling at work

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.
8) Consider that the rest of the samples not disregarded are samples from the marginal distributions f(x) and f(z).

4.2.2. *Why it works*

Consider a simple example in order to understand this intuitively: to obtain a posterior distribution of f(x, z) sampling from the conditionals f(x | z) and f(z | x). Figure 4.1 shows f(x,z) represented as lines of equal probability (as level curves in a map).

Take an arbitrary value for z, say z=1. Sample a random number, which is the univariate 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 density of probability between a and b 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 function. Suppose that the number sampled is x = 4. Now sample a random number from the conditional density f(z | x=4). This function is represented in figure 4.1 as the line x=4, that has more density of probability between 'c' and 'd' 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 function. Suppose that the number sampled is z = 5. Now sample a random number from the conditional density f(x | z=5). This function is represented in figure 4.1 as the line z=5, that has more density of probability between 'e' and 'f' 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 function. Suppose that the number sampled is x = 2. Now sample a random number from the conditional density f(z | x=2). This function is represented in figure 4.1 as the line x=2, that has more density of probability between 'g' and 'h' 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 function. Suppose that the number sampled is z=6, we will carry on with the same procedure until we obtain a chain of samples of the desired length.

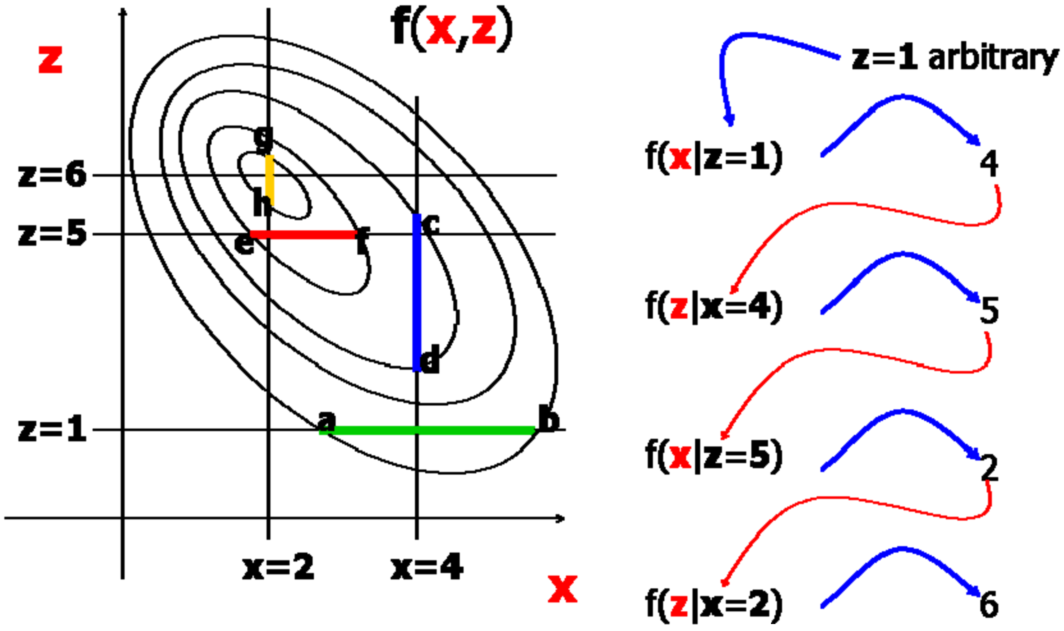

Figure 4.3. Gibbs sampling. The curves represent

Observe the tendency to sample from the highest areas of probability more often than to 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 find more samples in the highest areas of probability than in the lowest areas, thus we find random samples from the posterior distribution. This explains why the first points sampled should be discarded, and only after some cycles of iteration are samples taken at random.

4.2.3. *When it works*

1) *Strictly speaking, it cannot be demonstrated that we are finally 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 sampling from the posterior distribution begins*.

1) Even having a reducible chain and when the tests ensure convergence, the converged distribution may not be stationary. Sometimes there are large sequences of sampling that give the impression of stability, and after many iterations the chains moves 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 to 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 Monte-Carlo 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). Good practical textbooks dealing with MCMC application are Gelman et al. (2003), Gilks et al. (1996) and Robert and Casella (2004).

It should be noted that these difficulties are similar to finding a global maximum in multivariate likelihood with several fixed and random effects. With small databases maximum likelihood should not be used, since most properties of the method are asymptotic. With a large database, second derivative algorithms cannot be used by operative reasons, thus there is a formal incertitude about whether the maximum found is global or local. Here again, people use several starting values and examine the behaviour of their results. Complex models are difficult to handle in one or the other paradigm. However, MCMC techniques transform multivariate problems in univariate approaches, and inferences are made using probabilities, which has as 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 and they converge, we have the psychological persuasion that no convergence problems were found. We have no limit for the number of chains; in very simple problems, like linear models with 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 can be computed..

2. *Length of the chains*: It is customary to give 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.

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 inspection of the chain. In many problems this is not difficult, and in simple problems convergence is raised after little iteration. 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 the chain actually has, thus visual inspection is a common method to determine the burn-in.

4. *Sampling lag*: Samples are correlated. We have seen in 4.2.2 how we start in sampling in a part of the function and that it is not probable to jump to the other side of the posterior distribution for a new sample. 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 be taken into account.

Collecting a number of samples high enough is not a problem, we can arrive to the accuracy we desire, but to avoid collecting a high number of samples, consecutive samples are disregarded and, for example, only one each 20 samples is collected, which decreases the correlation between two consecutive samples. This is called the ‘lag’ between samples.

5. *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 samples highly correlated that will lead to the same precision as 35 uncorrelated samples. The actual sample size gives an idea about the real sample size we have, since to have a high number of samples highly correlated does not give much information.

In Tables of results

6. *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 values of the tests; I think this is rather irrelevant as far as they did not detected lack of convergence.

7. *Monte Carlo s.e.* (may be accompained by efective sample size and autocorrelation): 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 usually in two ways, using groups of samples and examining the sample means, or using temporal series techniques. Current software for MCMC usually gives the MCse. We should augment the sample until this error becomes irrelevant (for example, when it is 10 times lower than the standard deviation of the posterior distribution).

8. *Point estimates*: Median, mean, mode. When distributions are approximately symmetrical we should give one of them. I prefer the median for reasons explained in this book, but the mean and the mode are more frequently given.

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

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

11. *Probabilities*: P(d)>0, P(d>Relevant), Probability of similarity.