Chapter 7
Gibbs Sampler

Our ultimate objective is to simulate a gaussian random function with a specified covariance structure, given the observed lithotypes (facies) at sample points. As the lithotypes are known at these points, the corresponding gaussian variables must lie in certain intervals or sets but their values are not known. The difficulty is that these random functions conditioned to the constraints are no longer gaussian random functions.

In Chap. 2, the two step procedure used was presented from a theoretical point of view. Here we give a "maths-lite" presentation to illustrate the key concepts. The first section presents several examples to explain why we have recourse to a Gibbs sampler to generate gaussian values at sample points that have the right covariance and belong to the right intervals. Once we have this set of point values, any method for conditionally simulating gaussian random functions can be used; for example, turning bands together with a conditioning kriging, sequential gaussian simulations, LU decomposition, etc. See Chilès and Delfiner (1999), Lantuéjoul (2002a, b) or Deutch and Journel (1992). As these techniques are well known, we will not dwell on them here.

Why We Need a Two Step Simulation Procedure

The aim of this section is to highlight the difficulties of simulating gaussian random functions subject to interval constraints. To do this we consider three simple cases where there are only two points:

- With no constraints
- With interval constraints on one variable
- With interval constraints on both variables

In the first case, the conditional distribution of Z(x) given Z(y) turns out to be a gaussian distribution but this is no longer true in the other two cases. The conditional distributions are merely proportional to gaussians.

M. Armstrong et al., *Plurigaussian Simulations in Geosciences*,
DOI 10.1007/978-3-642-19607-2_7,

Simulating Z(x) and Z(y) When There Are No Constraints

Consider two gaussian variables Z(x) and Z(y) with a correlation coefficient, ρ. In order to simulate Z(x) given Z(y) we need to know its conditional distribution which can be deduced from the joint distribution of the two variables:

$$g(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{(u^2+v^2-2\rho uv)}{2(1-\rho^2)}\right\}$$

where u and v represent z(x) and z(y) respectively. This can be rewritten as

$$g(u,v) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(u-\rho v)^2}{2\,\sigma^2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-v^2}{2}\right\} \tag{7.1}$$

where $\sigma^2 = 1-\rho^2$. The second term is just the marginal distribution of Z(y). The first is the conditional distribution that we are looking for. It is clearly a gaussian distribution with mean, ρv, and variance σ^2. Equation (7.1) can be written as

$$g(u,v) = g_v(u)g(v) \tag{7.2}$$

This is equivalent to the well-known decomposition:

$$Z(x) = \rho Z(y) + \sigma R(x)$$

where R(x) is a N(0,1) residual that is independent of Z(y). If we estimate Z(x) given z(y), the simple kriging weight equals ρ, and the SK variance is $\sigma^2 = 1-\rho^2$.

To simulate pairs of values of Z(x) and Z(y), we first draw two independent N(0,1) values for Z(y) and R(x), then we substitute them into the decomposition formula to get Z(x). Alternatively we could say that we draw one realisation v of a N(0,1) variable for Z(y), followed a N(ρv,σ^2) variable for Z(x). In that case, we use the marginal distribution of Z(y) to draw a realisation of it, then the conditional distribution of Z(x) given Z(y) to draw the other value directly. This is only possible because the form of the conditional distribution is so simple.

Simulating Z(x) and Z(y) When Z(y) Belongs to an Interval

In this case Z(y) is known to lie in a specified interval, I, and we want to simulate the pair of variables, Z(x) and Z(y), given that Z(y) lies in that interval. The joint density of the two variables is now

$$h(u,v) = k\, g(u,v) 1_I(v)$$

where $1_I(v)$ is the indicator function for the interval I and k is the normation factor required to ensure that the integral of h(u,v) sums to 1. So the joint density is

$$h(u,v) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\frac{(u-\rho v)^2}{\sigma^2}\right\} \times \frac{k}{\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\frac{v^2}{}\right\} 1_I(v)$$

This can be written as

$$h(u,v) = h_V(u)\, h(v)$$

It is clear that h(v) is the marginal distribution of a gaussian variable Z(y) restricted to the interval I. To show that it is the marginal distribution, we just have to prove that

$$\int_{\Re} h_v(u)\, du = 1$$

Integrating h(u, v) with respect to u gives

$$\int_{\Re} h(u,v)\, du = \int_{\Re} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(u-\rho v)^2}{2\sigma^2}\right\} du \quad \frac{k}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} 1_I(v)$$

As the first term on the right hand side is just the integral of a $N(\rho v,\sigma^2)$ variable, it equals 1, which gives us the required result. In order to simulate these we use the same interpretation as before: first simulate Z(y) in the interval I then simulate Z(x) given that $Z(y) \in I$. The first simulation is just a truncated gaussian; the second one corresponds to simulating an independent $N(\rho v,\sigma^2)$ variable. That is, we are still using the classical decomposition.

But there is a fundamental change in the marginal distribution of Z(x). To see this, we integrate the joint density with respect to v using (7.2) written as $g(u,v) = g_u(v)g(u)$:

$$\begin{aligned}\int_{\Re} h(u,v)\, dv &= \int_{\Re} k\, g(u,v)\ \ 1_I(v)\, dv \\ &= \int_{\Re} \frac{1}{\sigma\sqrt{2\pi}}\ 1_I(v)\ \exp\left\{-\frac{(v-\rho u)^2}{2\sigma^2}\right\} dv \times \ \frac{k}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\}\end{aligned}$$

If we let $t = v - \rho u$, then $v \in I \Leftrightarrow t \in I - \rho u$ and the first term on the right becomes

$$\int_{\Re} \frac{1}{\sigma\sqrt{2\pi}}\, 1_{I-\rho u}(t) \exp\left\{-\frac{t^2}{2\sigma^2}\right\} dt = E\left[1_{I-\rho u}\right]$$

As $E[1_{I-\rho u}] = P_\sigma[I - \rho u]$

$$\int_{\Re} h(u, v)dv = k\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-u^2}{2}\right\} \times P_\sigma(I - \rho u)$$

This shows that the marginal density of Z(x) given that Z(y) $\in$ I, is no longer gaussian. It is proportional to a gaussian density but is multiplied by the probability that the first variable lies in the interval, $I - \rho u$. To illustrate the impact of this change, we have plotted this distribution for two intervals: $[-0.5, 0.5]$ and $[2, 3]$, and for two different correlation factors, -0.5 and 0.8. Figures 7.1 and 7.2 present the resulting curves.

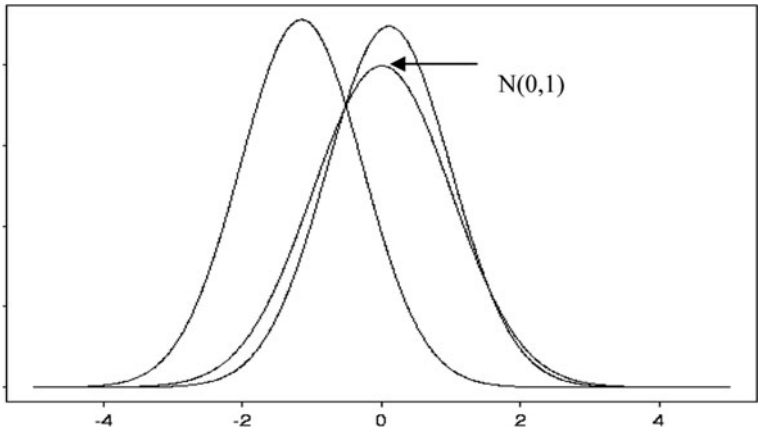

Fig. 7.1 Probability densities functions for the marginal distributions for the two intervals, for the case where $\rho = -0.5$ together with the N(0,1) density for comparison purposes

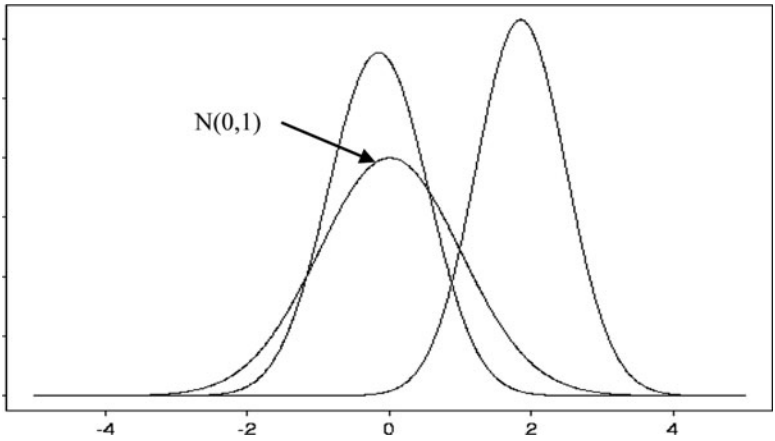

Fig. 7.2 Probability densities functions for the marginal distributions for the two intervals, for the case where $\rho = 0.8$ together with the N(0,1) density for comparison purposes

Simulating Z(x) and Z(y) When Both Belong to Intervals

Now suppose that Y(x) belongs to I_1 and Y(y) belongs to I_2. Their joint density is

$$h(u, v) = kg(u, v)\, 1_{I_1}(u)\, 1_{I_2}(v)$$

where k is the appropriate normation factor.

$$h(u, v) = \frac{k}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(u - \rho v^2)}{2\sigma^2}\right\} 1_{I_1}(u) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} 1_{I_2}(v)$$

As expected, the marginal distributions are no longer gaussian or even truncated gaussian. Integrating with respect to u gives

$$\int_{\Re} h(u, v)\, du \propto \frac{1}{\sqrt{2\pi}} P_{\sigma_2}(I_1 - \rho v)\, \exp\left\{\frac{-v^2}{2}\right\} 1_{I_2}(v) \tag{7.3}$$

Similarly integrating with respect to v gives

$$\int_{\Re} h(u, v)\, dv \propto \frac{1}{\sqrt{2\pi}} P_{\sigma_1}(I_2 - \rho u)\, \exp\left\{\frac{-u^2}{2}\right\} 1_{I_1}(u) \tag{7.4}$$

Because of the increasing difficult in simulating these distributions directly as the number of interval constraints increases, we rapidly reach the point where direct simulation is no longer practicable and we have to resort to an indirect approach such as the Gibbs sampler.

These examples suggest the idea of using a two step procedure to conditionally simulate gaussian values when some of them are constrained to lie in specified intervals. The two steps are:

1. Generating gaussian values at data points, in the prescribed intervals and with the right covariance structure
2. Using any algorithm for conditionally simulating gaussian random functions given the values generated in step (1)

Direct Simulation Using an Acceptance/Rejection Procedure

Having seen that the first step is to generate a set of gaussian values at sample points, the next question is how this should be done. We might be tempted to try an acceptance/rejection procedure, by generating gaussian values with the right covariance structure and rejecting those lying outside the specified intervals. If only 10 or 20 samples were available this could be done directly using an LU decomposition

(i.e. a Cholesky decomposition of the covariance matrix). The LU decomposition could be carried out for much larger matrices (up to about 500 × 500). The limiting factor in the procedure is the rate of rejection. For example, suppose that ten samples were available and that there were only two facies each present 50% of the time. Then if there was no spatial correlation (pure nugget effect), the probability of getting all ten values in the right intervals would be 1 in 2^{10}; that is, about 1 in 1,000. This procedure becomes prohibitively slow as the number of samples increases. As there are usually hundreds or thousands of data in mining and petroleum applications, another approach is needed. This is why we have to resort to more complicated methods.

Gibbs Sampler

Statisticians routinely use iterative methods based on Markov chain Monte Carlo simulations (MCMC, for short) for sampling complicated distributions and for estimating parameter values. The best known are the Hastings-Metropolis algorithm and the Gibbs sampler. The latter is a particular case of the Hastings-Metropolis method. See Meyn and Tweedie (1993), Cowles and Carlin (1996) and Robert (1996) for information on these methods. Freulon (1992) and Freulon and de Fouquet (1993) adapted the Gibbs sampler to truncated gaussian simulations. To introduce this technique, we present an example to show how this method works and to illustrate the concept of convergence.

Four Sample Example

Suppose that there are only two lithotypes F_1 and F_2, and that they are present in equal proportions. So it is natural to use a zero threshold to separate them. Negative

Fig. 7.3 Simplified well or drill hole containing four samples

gaussian values correspond to F_1; positive ones, to F_2. Figure 7.3 shows a simplified well or drill hole containing four samples, with the top two belonging to F_1 and the other two belonging to F_2. The gaussian values assigned to the top samples must be negative, the others have to be positive.

The procedure relies on the standard decomposition of the gaussian random function into its simple kriging estimate and an orthogonal residual.

$$Z(x) = Z_{SK}(x) + \sigma_{SK} R(x)$$

where the index SK denotes the simple kriging estimate or its variance and R(x) is a N(0,1) residual. Note that for gaussian random variables, the SK estimate equals the conditional expectation.

Exponential Variogram

Suppose that the four samples are 1m apart and that the underlying variogram for the gaussians is an exponential with a unit sill and a scale parameter a $= 2$ m (i.e. a practical range of 6 m). The SK weights for estimating the top point using the other three points as data are:

$$\lambda_2 = 0.61, \quad \lambda_3 = 0, \quad \lambda_4 = 0 \quad \text{and} \quad \sigma^2_{SK} = 0.63 \Rightarrow \sigma_{SK} = 0.79$$

(Points are numbered from the top down). By symmetry the weights are the same but in the reverse order when kriging the fourth point from the other three. The weights for the second point (or similarly the third one) are:

$$\lambda_1 = 0.44, \quad \lambda_3 = 0.44, \quad \lambda_4 = 0 \quad \text{and} \quad \sigma^2_{SK} = 0.46 \Rightarrow \sigma_{SK} = 0.68$$

As the configuration does not change from one iteration to the next the weights remain the same.

Step 1: Initialising the Procedure

The first step consists of choosing gaussian values that belong to the appropriate intervals. Here we select $(-1, -1, +1, +1)$

Step 2: Iterative Procedure

Simple kriging is applied to the points in turn. For example, the value of the top point is kriged using the other three points as input data. Then we move down to the second point and krige it using the initial values for the points below it and the new

updated value for the top point. After completing the second point we move down to the third one which is kriged using the updated values for the points above it and the old value for the point below it. Similarly for the fourth point. When all the points have been updated by kriging, one iteration has been completed.

Point No 1

The kriged estimate for $Z(x_1)$, abbreviated to Z(1), based on the initial values (i.e. −1, +1, +1) for the other three points is:

$$Z_{SK}(1) = -1 \times 0.61 + 1 \times 0 + 1 \times 0 = -0.61$$

And the corresponding residual must satisfy

$$R(1) \leq -(-0.61)/0.79 = 0.77$$

Suppose for argument's sake that we draw a value of 0.52 (from a N(0,1) distribution). Then the updated value would be

$$Z(1) = -0.61 + 0.79 \times 0.52 = -0.20$$

Point No 2

The kriged estimate for Z(2) based on the initial values for Z(3) and Z(4), and the updated value of Z(1) is:

$$Z_{SK}(2) = -0.20 \times 0.44 + 1 \times 0.44 + 1 \times 0 = 0.42$$

The corresponding residual must satisfy

$$R(2) \leq 0.42/0.68 = 0.62$$

If we draw a value of −0.75, then the updated value of Z(2) is

$$Z(2) = 0.42 + 0.68 \times -0.75 = -0.09$$

Point No 3

Following the same procedure, the kriged estimate for Z(3) and the inequality to be satisfied by its residual are

$$Z_{SK}(3) = -0.20 \times 0 - 0.09 \times 0.44 + 1 \times 0.44 = +0.40$$

$$R(3) \geq 0.40/0.68 = 0.59$$

If we draw a value of 0.28, then the updated value of Z(3) is 0.21.

Point No 4

In the same way, the kriged estimate for Z(4) and the inequality to be satisfied by its residual are

$$Z_{SK}(4) = 0 \times (0.20) + 0 \times (-0.09) + 0.61 \times 0.21 = 0.13$$
$$R(4) \geq -0.13/0.79 = -0.16$$

Drawing a value of 0.63 would give an updated value of 0.63 for Z(4).

Results

Table 7.1 summarises the intermediate results during first iteration. This updating procedure is repeated iteratively, in general for several hundred or several thousand iterations. Table 7.2 shows the results of the first five iterations.

Alternative Updating Strategies

In the previous example, individual points were sequentially updated. A variant of this consists of sequentially updating from the top down, then from the bottom up on the next iteration.

Table 7.1 Successive steps in the first iteration of this Gibbs sampler

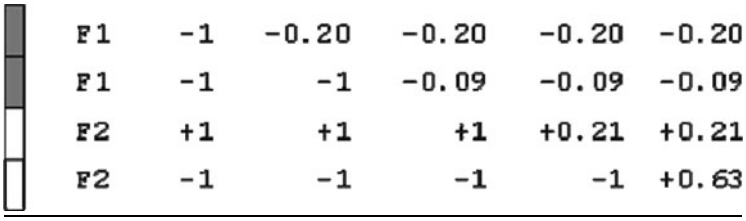

Table 7.2 Results of first five iterations of the Gibbs sampler

Pt N°	Initial	No 1	No2	No3	N° 4	N° 5
1	−1	−0.20	−0.37	−1.34	−0.34	−0.13
2	−1	−0.09	−0.35	−0.24	−0.31	−0.20
3	+1	+0.21	+0.15	+0.05	+0.19	+0.20
4	−1	+0.63	+0.86	+0.10	+0.26	+0.32

Blocking Factor

It is also possible to update blocks of points simultaneously. For example, the four points could be grouped into two blocks each consisting of two points. There are three possible groupings of this type:

- Points 1 & 2 and Points 3 & 4
- Points 1 & 3 and Points 2 & 4
- Points 1 & 4 and Points 2 & 3

In the first case, the values of the top two points are updated using the values of the other two as the conditioning data (i.e. using kriging) and simulated in the right interval with the right correlations, and vice versa for the other pair. We will illustrate this procedure later in the chapter. As was shown in Chap. 2 suitably chosen blocking strategies can significantly improve the speed of convergence.

Experimentally Testing Convergence

Having seen how the procedure works, several questions need to be answered. Firstly, does the algorithm converge? If so, after how many iterations? What factors affect the speed of convergence? How should we choose the initial values?

Burn-in Period

In this section we illustrate the difference between the initial burn-in period and the subsequent stationary part of the Markov chain. To do this we continue the previous example but the variogram is changed to a gaussian model with a practical range of 3. So the correlation between adjoining samples is 0.95. Five hundred iterations of the corresponding Gibbs sampler were run starting from a very extreme set of initial values (+5,+5,−5,−5). This choice lengthens the initial burn-in period, making it visually much more obvious.

Figure 7.4 shows the output for each component as a function of the number of iterations. The values of the first component (top left) decrease steadily from the initial value of +5 until they are below 1.0. The curve seems to stabilise after approximately 100 iterations so the burn-in period must be at least this long. Similarly for the third and fourth components. But it appears to be much shorter for the other component (bottom left), about 20 iterations. This shows that the burn-in period need not be the same for all components in a Gibbs sampler. In MCMC theory it is well-known that different states can have different rates of convergence; see Meyn and Tweedie (1993, pp 362–363).

The implications of not necessarily having the same burn-in period for all components are important in practice. When there are only four components it is

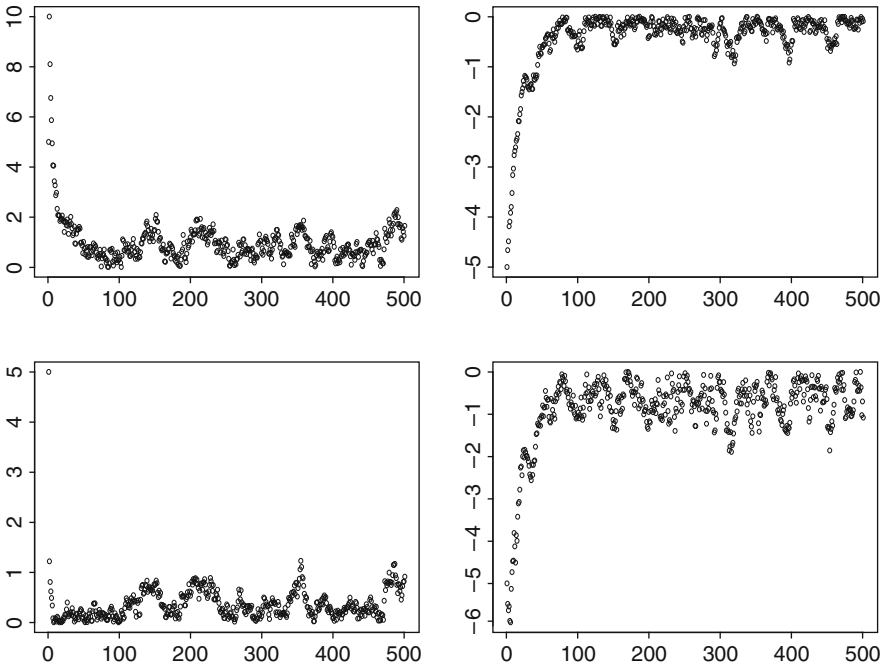

Fig. 7.4 Output of the Gibbs sampler for all 4 components for 500 iterations, starting from initial values of (+5,+5,−5,−5). The second component (*lower left*) seems to stabilise after about 20 iterations components whereas the other three are much slower. They take at least 100 iterations to reach their stationary distribution

possible to check the convergence of all of them but if there were 1,000 samples it would be virtually impossible to inspect the output of the Gibbs sampler for all 1,000 components. We could only check a few of them visually and we might have the bad luck to choose those with shorter burn-ins. Inspecting the results for selected components gives us an idea of the burn-in period but it is not foolproof.

Effect of the Range on the Burn-in Period

Several factors including the range of the variogram and the number of components have a marked effect on the length of the burn-in period. To illustrate the effect of the range, we repeated the previous example using a practical range of 5 instead of 3. This increases the correlation between adjoining samples from 0.95 to 0.98. Figure 7.5 shows the output.

Whereas the burn-in period for the first component was about 100 beforehand, it is now closer to 500. Conversely decreasing the range would decrease the burn-in period. Looking at Fig. 7.5 we also notice how smooth the curves are

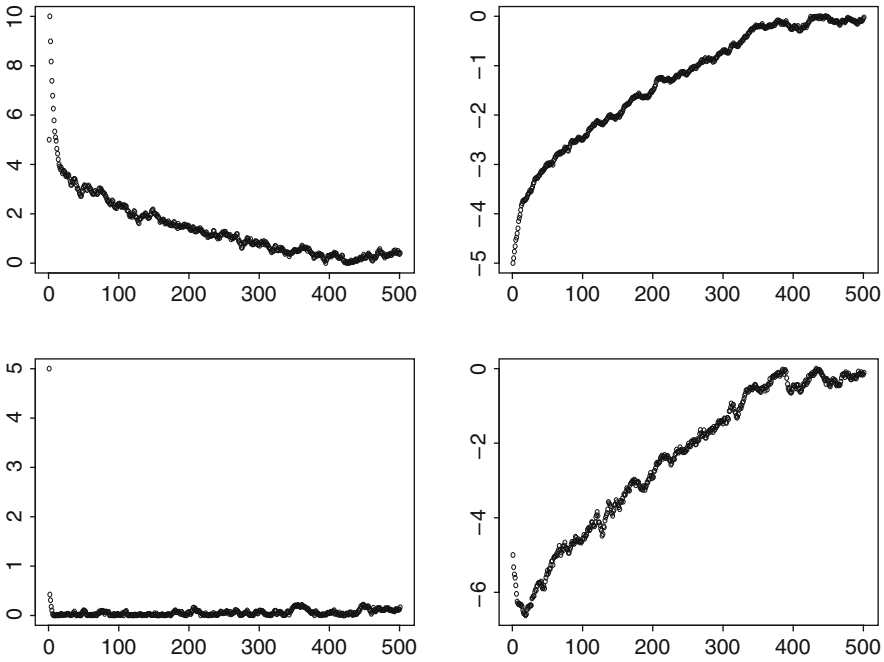

Fig. 7.5 Output of the Gibbs sampler for the first of 4 components for 500 iterations, starting from initial values of (+5,+5,−5,−5). Compared to Fig. 7.4, the correlation between adjoining points has been increased from 0.95 to 0.98. Note the increase in the burn-in period

compared to the corresponding ones in Fig. 7.4. The strong serial correlation between successive values makes it more difficult to determine whether the Markov chain has converged.

These two examples show that it is not simple to judge whether a Gibbs sampler has reached its stationary distribution just by studying the output from a single run (even a very long one). It would be better to run a large number of samplers in parallel and study their output after 1, 5, 10, . . . , 50 iterations and so on. Ideally we should compare the experimental distribution of the output with the stationary distribution. How could this be tested experimentally?

A multivariate normal distribution is fully specified when the means and the covariance matrix are known. By extension, a truncated gaussian is fully defined when the truncation thresholds are known together with the means and the covariance matrix for the full distribution. Having said that, it is clear that after truncation the means are not going to be the same as before. See (7.3) and (7.4). Nor are the variances or the correlations. In the case under study, the four components come from a quadrivariate normal distribution with the covariance matrix given in Table 7.3.

Because of the high correlation between adjoining samples, its density is an elongated cigar shape. Figure 7.6 shows a diagram representing the bivariate densities of X_1 and X_2, and X_2 and X_3 respectively. The impact of different types

Table 7.3 Quadrivariate covariance matrix for the 4-sample case

$$\begin{bmatrix} 1 & 0.95 & 0.80 & 0.61 \\ 0.95 & 1 & 0.95 & 0.80 \\ 0.80 & 0.95 & 1 & 0.95 \\ 0.61 & 0.80 & 0.95 & 1 \end{bmatrix}$$

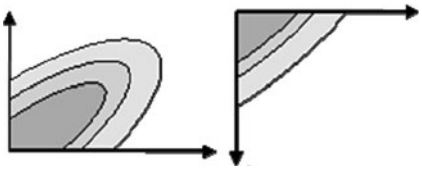

Fig. 7.6 Schematic representation of the bivariate densities of X_1 and X_2 (*left*) and of X_2 and X_3 (*right*)

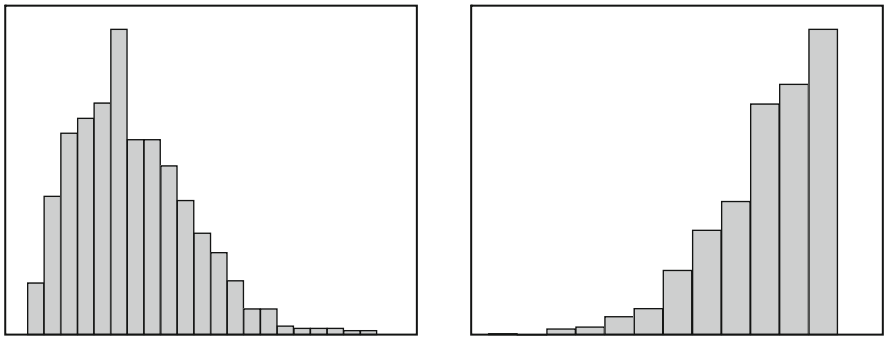

Fig. 7.7 The marginal distributions of X_1 and of X_3

of truncations is evident from these. In one case, we are dealing with the elongated part of the ellipse whereas in the other, it is merely a triangular "corner". Intuition can often be misleading when trying to guess the properties of truncated gaussian distributions. For example many people expect the marginal distributions in this example to be "half gaussians". The marginal distributions given in Fig. 7.7 show just how wrong intuition can be and also confirms what was shown in the two variable case given at the beginning of the chapter.

The Impact of Different Parallel Runs

Up till now we have illustrated the difference between the burn-in period and the stationary part by focussing on individual components. An alternative is to run

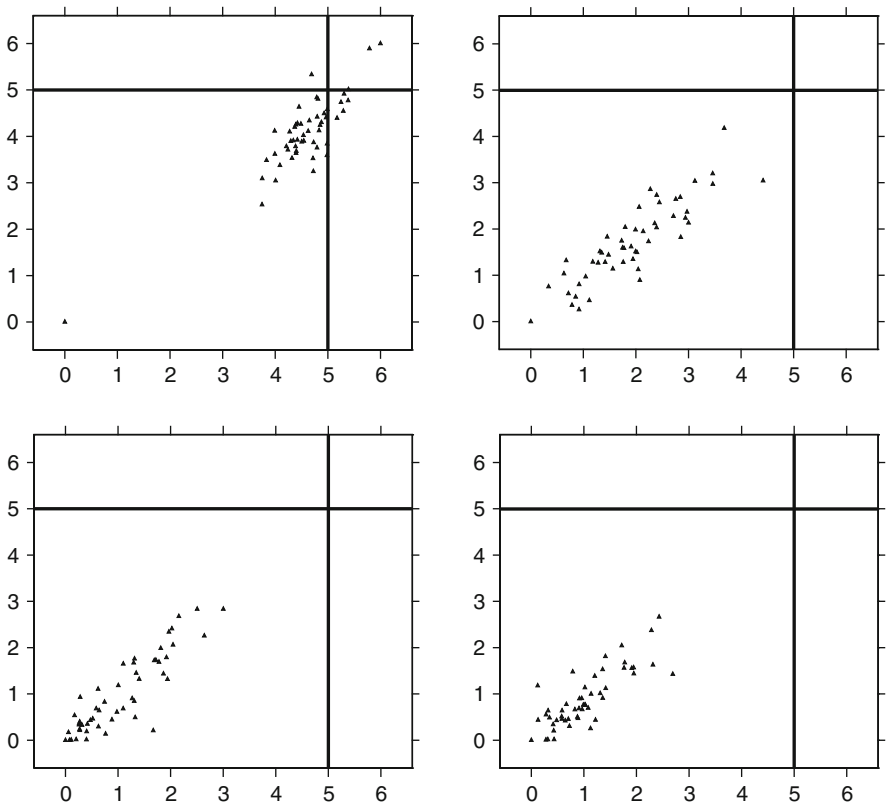

Fig. 7.8 The values of the first and second components after 1 iteration (*top left*), then 5 (*top right*), 10 (*bottom left*) and 50 iterations (*bottom right*) starting out from initial values of +5 (as indicated by the crosshairs)

many Gibbs samplers in parallel and study the results after a certain number of iterations. Figure 7.8 plots the first and second components for 50 parallel runs. Figure 7.8a shows their locations after a single iteration; both started out from an initial value of +5. Figures 7.8b–d give the output after 5, 10 and 50 iterations. As expected, the centre of the cloud moves downwards and disperses outward from this. Initially the distribution is far from the target cloud but as the number of iterations increases, it steadily tends toward it.