Chapter 4

# The Gibbs Sampler

## 4.1 Introduction

In section 3.3 we have seen that, using importance sampling, we can approximate an expectation $\mathbb{E}_f(h(X))$ without having to sample directly from $f$. However, finding an instrumental distribution which allows us to *efficiently* estimate $\mathbb{E}_f(h(X))$ can be difficult, especially in large dimensions.

In this chapter and the following chapters we will use a somewhat different approach. We will discuss methods that allow obtaining an *approximate* sample from $f$ without having to sample from $f$ directly. More mathematically speaking, we will discuss methods which generate a Markov chain whose stationary distribution is the distribution of interest $f$. Such methods are often referred to as Markov Chain Monte Carlo (MCMC) methods.

*Example 4.1 (Poisson change point model).* Assume the following Poisson model of two regimes for $n$ random variables $Y_1, \ldots, Y_n$.[1]

$$\begin{aligned} Y_i &\sim \mathsf{Poi}(\lambda_1) \quad \text{for} \quad i = 1, \ldots, M \\ Y_i &\sim \mathsf{Poi}(\lambda_2) \quad \text{for} \quad i = M+1, \ldots, n \end{aligned}$$

A suitable (conjugate) prior distribution for $\lambda_j$ is the $\mathsf{Gamma}(\alpha_j, \beta_j)$ distribution with density

$$f(\lambda_j) = \frac{1}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j - 1} \beta_j^{\alpha_j} \exp(-\beta_j \lambda_j).$$

The joint distribution of $Y_1, \ldots, Y_n$, $\lambda_1$, $\lambda_2$, and $M$ is

$$\begin{aligned} f(y_1, \ldots, y_n, \lambda_1, \lambda_2, M) &= \left( \prod_{i=1}^{M} \frac{\exp(-\lambda_1)\lambda_1^{y_i}}{y_i!} \right) \cdot \left( \prod_{i=M+1}^{n} \frac{\exp(-\lambda_2)\lambda_2^{y_i}}{y_i!} \right) \\ &\quad \cdot \frac{1}{\Gamma(\alpha_1)} \lambda_1^{\alpha_1 - 1} \beta_1^{\alpha_1} \exp(-\beta_1 \lambda_1) \cdot \frac{1}{\Gamma(\alpha_2)} \lambda_2^{\alpha_2 - 1} \beta_2^{\alpha_2} \exp(-\beta_2 \lambda_2). \end{aligned}$$

If $M$ is known, the posterior distribution of $\lambda_1$ has the density

$$f(\lambda_1 | Y_1, \ldots, Y_n, M) \propto \lambda_1^{\alpha_1 - 1 + \sum_{i=1}^{M} y_i} \exp(-(\beta_1 + M)\lambda_1),$$

so

$$\begin{aligned} \lambda_1 | Y_1, \ldots Y_n, M &\sim \mathsf{Gamma}\left( \alpha_1 + \sum_{i=1}^{M} y_i, \beta_1 + M \right) \qquad (4.1) \\ \lambda_2 | Y_1, \ldots Y_n, M &\sim \mathsf{Gamma}\left( \alpha_2 + \sum_{i=M+1}^{n} y_i, \beta_2 + n - M \right). \qquad (4.2) \end{aligned}$$

[1] The probability distribution function of the $\mathsf{Poi}(\lambda)$ distribution is $p(y) = \frac{\exp(-\lambda)\lambda^y}{y!}$.



Now assume that we do not know the change point $M$ and that we assume a uniform prior on the set $\{1, \ldots, M-1\}$. It is easy to compute the distribution of $M$ given the observations $Y_1, \ldots Y_n$, and $\lambda_1$ and $\lambda_2$. It is a discrete distribution with probability density function proportional to

$$p(M | Y_1, \ldots, Y_n, \lambda_1, \lambda_2) \propto \lambda_1^{\sum_{i=1}^{M} y_i} \cdot \lambda_2^{\sum_{i=M+1}^{n} y_i} \cdot \exp((\lambda_2 - \lambda_1) \cdot M) \qquad (4.3)$$

The conditional distributions in (4.1) to (4.3) are all easy to sample from. It is however rather difficult to sample from the joint posterior of $(\lambda_1, \lambda_2, M)$. ◁

The example above suggests the strategy of alternately sampling from the (full) conditional distributions ((4.1) to (4.3) in the example). This tentative strategy however raises some questions.

- Is the joint distribution uniquely specified by the conditional distributions?
- Sampling alternately from the conditional distributions yields a Markov chain: the newly proposed values only depend on the present values, not the past values. Will this approach yield a Markov chain with the correct invariant distribution? Will the Markov chain converge to the invariant distribution?

As we will see in sections 4.3 and 4.4, the answer to both questions is — under certain conditions — yes. The next section will however first of all state the Gibbs sampling algorithm.

## 4.2 Algorithm

The Gibbs sampler was first proposed by Geman and Geman (1984) and further developed by Gelfand and Smith (1990). Denote with $x_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$.

**Algorithm 4.1 ((Systematic sweep) Gibbs sampler).** Starting with $(X_1^{(0)}, \ldots, X_p^{(0)})$ iterate for $t = 1, 2, \ldots$

1. Draw $X_1^{(t)} \sim f_{X_1|X_{-1}}(\cdot | X_2^{(t-1)}, \ldots, X_p^{(t-1)})$.

…

j. Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot | X_1^{(t)}, \ldots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$.

…

p. Draw $X_p^{(t)} \sim f_{X_p|X_{-p}}(\cdot | X_1^{(t)}, \ldots, X_{p-1}^{(t)})$.

Figure 4.1 illustrates the Gibbs sampler. The conditional distributions as used in the Gibbs sampler are often referred to as *full conditionals*. Note that the Gibbs sampler is *not* reversible. Liu et al. (1995) proposed the following algorithm that yields a reversible chain.

**Algorithm 4.2 (Random sweep Gibbs sampler).** Starting with $(X_1^{(0)}, \ldots, X_p^{(0)})$ iterate for $t = 1, 2, \ldots$

1. Draw an index $j$ from a distribution on $\{1, \ldots, p\}$ (e.g. uniform)
2. Draw $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot | X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_p^{(t-1)})$, and set $X_\iota^{(t)} := X_\iota^{(t-1)}$ for all $\iota \neq j$.

## 4.3 The Hammersley-Clifford Theorem

An interesting property of the full conditionals, which the Gibbs sampler is based on, is that they fully specify the joint distribution, as Hammersley and Clifford proved in 1970[2]. Note that the set of marginal distributions does *not* have this property.

[2] Hammersley and Clifford actually never published this result, as they could not extend the theorem to the case of non-positivity.

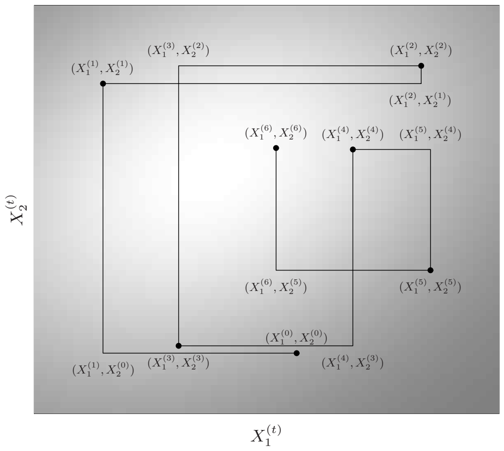



**Figure 4.1.** Illustration of the Gibbs sampler for a two-dimensional distribution

**Definition 4.1 (Positivity condition).** *A distribution with density $f(x_1, \ldots, x_p)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if $f_{X_i}(x_i) > 0$ for all $x_1, \ldots, x_p$ implies that $f(x_1, \ldots, x_p) > 0$.*

The positivity condition thus implies that the support of the joint density $f$ is the Cartesian product of the support of the marginals $f_{X_i}$.

**Theorem 4.2 (Hammersley-Clifford).** *Let $(X_1, \ldots, X_p)$ satisfy the positivity condition and have joint density $f(x_1, \ldots, x_p)$. Then for all $(\xi_1, \ldots, \xi_p) \in supp(f)$*

$$f(x_1, \ldots, x_p) \propto \prod_{j=1}^{p} \frac{f_{X_j|X_{-j}}(x_j|x_1, \ldots, x_{j-1}, \xi_{j+1}, \ldots, \xi_p)}{f_{X_j|X_{-j}}(\xi_j|x_1, \ldots, x_{j-1}, \xi_{j+1}, \ldots, \xi_p)}$$

*Proof.* We have

$$f(x_1, \ldots, x_{p-1}, x_p) = f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1}) f(x_1, \ldots, x_{p-1}) \tag{4.4}$$

and by complete analogy

$$f(x_1, \ldots, x_{p-1}, \xi_p) = f_{X_p|X_{-p}}(\xi_p|x_1, \ldots, x_{p-1}) f(x_1, \ldots, x_{p-1}), \tag{4.5}$$

thus

$$\begin{aligned} f(x_1, \ldots, x_p) &\overset{(4.4)}{=} \underbrace{f(x_1, \ldots, x_{p-1})}_{\overset{(4.5)}{=} f(x_1, \ldots, x_{p-1}, \xi_p)/f_{X_p|X_{-p}}(\xi_p|x_1, \ldots, x_{p-1})} f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1}) \\ &= f(x_1, \ldots, x_{p-1}, \xi_p) \frac{f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1})}{f_{X_p|X_{-p}}(\xi_p|x_1, \ldots, x_{p-1})} \\ &= \ldots \\ &= f(\xi_1, \ldots, \xi_p) \frac{f_{X_1|X_{-1}}(x_1|\xi_2, \ldots, \xi_p)}{f_{X_1|X_{-1}}(\xi_1|\xi_2, \ldots, \xi_p)} \cdots \frac{f_{X_p|X_{-p}}(x_p|x_1, \ldots, x_{p-1})}{f_{X_p|X_{-p}}(\xi_p|x_1, \ldots, x_{p-1})} \end{aligned}$$

The positivity condition guarantees that the conditional densities are non-zero. □

Note that the Hammersley-Clifford theorem does *not* guarantee the existence of a joint probability distribution for every choice of conditionals, as the following example shows. In Bayesian modeling such problems mostly arise when using improper prior distributions.



*Example 4.2.* Consider the following "model"

$$\begin{aligned} X_1|X_2 &\sim \mathsf{Expo}(\lambda X_2) \\ X_2|X_1 &\sim \mathsf{Expo}(\lambda X_1), \end{aligned}$$

for which it would be easy to design a Gibbs sampler. Trying to apply the Hammersley-Clifford theorem, we obtain

$$f(x_1, x_2) \propto \frac{f_{X_1|X_2}(x_1|\xi_2) \cdot f_{X_2|X_1}(x_2|x_1)}{f_{X_1|X_2}(\xi_1|\xi_2) \cdot f_{X_2|X_1}(\xi_2|x_1)} = \frac{\lambda\xi_2 \exp(-\lambda x_1 \xi_2) \cdot \lambda x_1 \exp(-\lambda x_1 x_2)}{\lambda\xi_2 \exp(-\lambda \xi_1 \xi_2) \cdot \lambda x_1 \exp(-\lambda x_1 \xi_2)} \propto \exp(-\lambda x_1 x_2)$$

The integral $\int \int \exp(-\lambda x_1 x_2) \, dx_1 \, dx_2$ however is not finite, thus there is no two-dimensional probability distribution with $f(x_1, x_2)$ as its density. ◁

## 4.4 Convergence of the Gibbs sampler

First of all we have to analyse whether the joint distribution $f(x_1, \ldots, x_p)$ is indeed the stationary distribution of the Markov chain generated by the Gibbs sampler[3]. For this we first have to determine the transition kernel corresponding to the Gibbs sampler.

**Lemma 4.3.** *The transition kernel of the Gibbs sampler is*

$$\begin{aligned} K(\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) &= f_{X_1|X_{-1}}(x_1^{(t)}|x_2^{(t-1)}, \ldots, x_p^{(t-1)}) \cdot f_{X_2|X_{-2}}(x_2^{(t)}|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)}) \cdot \ldots \\ &\quad \cdot f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)}, \ldots, x_{p-1}^{(t)}) \end{aligned}$$

*Proof.* We have

$$\begin{aligned} &\mathbb{P}(\mathbf{X}^{(t)} \in \mathcal{X}|\mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) = \int_{\mathcal{X}} f_{(\mathbf{X}^t|\mathbf{X}^{(t-1)})}(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) \, d\mathbf{x}^{(t)} \\ &= \int_{\mathcal{X}} \underbrace{f_{X_1|X_{-1}}(x_1^{(t)}|x_2^{(t-1)}, \ldots, x_p^{(t-1)})}_{\text{corresponds to step 1. of the algorithm}} \cdot \underbrace{f_{X_2|X_{-2}}(x_2^{(t)}|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)})}_{\text{corresponds to step 2. of the algorithm}} \cdot \ldots \\ &\quad \cdot \underbrace{f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)}, \ldots, x_{p-1}^{(t)})}_{\text{corresponds to step p. of the algorithm}} \, d\mathbf{x}^{(t)} \end{aligned}$$

□

[3] All the results in this section will be derived for the systematic scan Gibbs sampler (algorithm 4.1). Very similar results hold for the random scan Gibbs sampler (algorithm 4.2).

**Proposition 4.4.** *The joint distribution $f(x_1,\ldots,x_p)$ is indeed the invariant distribution of the Markov chain $(\mathbf{X}^{(0)},\mathbf{X}^{(1)},\ldots)$ generated by the Gibbs sampler.*

*Proof.*

$$\begin{aligned}
&\int f(\mathbf{x}^{(t-1)})K(\mathbf{x}^{(t-1)},\mathbf{x}^{(t)})\,d\mathbf{x}^{(t-1)}\\
=&\int\cdots\underbrace{\underbrace{\int f(x_1^{(t-1)},\ldots,x_p^{(t-1)})\,dx_1^{(t-1)}}_{=f(x_2^{(t-1)},\ldots,x_p^{(t-1)})}f_{X_1|X_{-1}}(x_1^{(t)}|x_2^{(t-1)},\ldots,x_p^{(t-1)})}_{=f(x_1^{(t)},x_2^{(t-1)},\ldots,x_p^{(t-1)})}\cdots f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)},\ldots,x_{p-1}^{(t)})dx_2^{(t-1)}\ldots dx_p^{(t-1)}\\
=&\int\cdots\underbrace{\underbrace{\int f(x_1^{(t)},x_2^{(t-1)},\ldots,x_p^{(t-1)})\,dx_2^{(t-1)}}_{=f(x_1^{(t)},x_3^{(t-1)},\ldots,x_p^{(t-1)})}f_{X_2|X_{-2}}(x_2^{(t)}|x_1^{(t)},x_3^{(t-1)},\ldots,x_p^{(t-1)})}_{=f(x_1^{(t)},x_2^{(t)},x_3^{(t-1)},\ldots,x_p^{(t-1)})}\cdots f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)},\ldots,x_{p-1}^{(t)})dx_3^{(t-1)}\ldots dx_p^{(t-1)}\\
=&\ \ldots\\
=&\underbrace{\underbrace{\int f(x_1^{(t)},\ldots,x_{p-1}^{(t)},x_p^{(t-1)})\,dx_p^{(t-1)}}_{=f(x_1^{(t)},\ldots,x_{p-1}^{(t)})}f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)},\ldots,x_{p-1}^{(t)})}_{=f(x_1^{(t)},\ldots,x_p^{(t)})}\\
=&\ f(x_1^{(t)},\ldots,x_p^{(t)})
\end{aligned}$$

Thus according to definition 1.27 $f$ is indeed the invariant distribution. □





So far we have established that $f$ is indeed the invariant distribution of the Gibbs sampler. Next, we have to analyse under which conditions the Markov chain generated by the Gibbs sampler will converge to $f$.

First of all we have to study under which conditions the resulting Markov chain is irreducible[4]. The following example shows that this does not need to be the case.

*Example 4.3 (Reducible Gibbs sampler).* Consider Gibbs sampling from the uniform distribution on $C_1 \cup C_2$ with $C_1 := \{(x_1,x_2) : \|(x_1,x_2)-(1,1)\| \le 1\}$ and $C_2 := \{(x_1,x_2) : \|(x_1,x_2)-(-1,-1)\| \le 1\}$, i.e.

$$f(x_1,x_2) = \frac{1}{2\pi}\mathbb{I}_{C_1\cup C_2}(x_1,x_2)$$

Figure 4.2 shows the density as well the first few samples obtained by starting a Gibbs sampler with $X_1^{(0)} < 0$ and $X_2^{(0)} < 0$. It is easy to that when the Gibbs sampler is started in $C_2$ it will stay there and never reach $C_1$. The reason

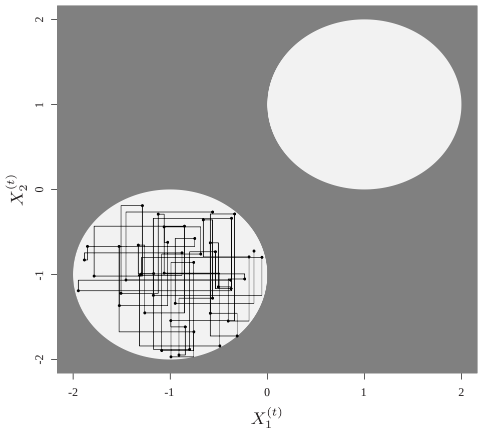



**Figure 4.2.** Illustration of a Gibbs sampler failing to sample from a distribution with unconnected support (uniform distribution on $\{(x_1,x_2) : \|(x_1,x_2)-(1,1)\| \le 1|\} \cup \{(x_1,x_2) : \|(x_1,x_2)-(-1,-1)\| \le 1|\}$)

for this is that the conditional distribution $X_2|X_1$ ($X_1|X_2$) is for $X_1 < 0$ ($X_2 < 0$) entirely concentrated on $C_2$. ◁

The following proposition gives a sufficient condition for irreducibility (and thus the recurrence) of the Markov chain generated by the Gibbs sampler. There are less strict conditions for the irreducibility and aperiodicity of the Markov chain generated by the Gibbs sampler (see e.g. Robert and Casella, 2004, Lemma 10.11).

**Proposition 4.5.** *If the joint distribution $f(x_1,\ldots,x_p)$ satisfies the positivity condition, the Gibbs sampler yields an irreducible, recurrent Markov chain.*

*Proof.* Let $\mathcal{X} \subset \text{supp}(f)$ be a set with $\int_{\mathcal{X}} f(x_1^{(t)},\ldots,x_p^{(t)})d(x_1^{(t)},\ldots,x_p^{(t)}) > 0$.

$$\int_{\mathcal{X}} K(\mathbf{x}^{(t-1)},\mathbf{x}^{(t)})d\mathbf{x}^{(t)} = \int_{\mathcal{X}} \underbrace{f_{X_1|X_{-1}}(x_1^{(t)}|x_2^{(t-1)},\ldots,x_p^{(t-1)})}_{>0\text{ (on a set of non-zero measure)}}\cdots\underbrace{f_{X_p|X_{-p}}(x_p^{(t)}|x_1^{(t)},\ldots,x_{p-1}^{(t)})}_{>0\text{ (on a set of non-zero measure)}}d\mathbf{x}^{(t)} > 0,$$

where the conditional densities are non-zero by the positivity condition. Thus the Markov Chain $(\mathbf{X}^{(t)})_t$ is strongly $f$-irreducible. As $f$ is the unique invariant distribution of the Markov chain, it is as well recurrent (proposition 1.28). □

[4] Here and in the following we understand by "irreducibilty" irreducibility with respect to the target distribution $f$.

If the transition kernel is absolutely continuous with respect to the dominating measure, then recurrence even implies Harris recurrence (see e.g. Robert and Casella, 2004, Lemma 10.9).

Now we have established all the necessary ingredients to state an ergodic theorem for the Gibbs sampler, which is a direct consequence of theorem 1.30.

**Theorem 4.6.** *If the Markov chain generated by the Gibbs sampler is irreducible and recurrent (which is e.g. the case when the positivity condition holds), then for any integrable function* $h : E \to \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} h(\mathbf{X}^{(t)}) \to \mathbb{E}_f\left(h(\mathbf{X})\right)$$

*for almost every starting value* $\mathbf{X}^{(0)}$. *If the chain is Harris recurrent, then the above result holds for every starting value* $\mathbf{X}^{(0)}$.

Theorem 4.6 guarantees that we can approximate expectations $\mathbb{E}_f\left(h(\mathbf{X})\right)$ by their empirical counterparts using *a single* Markov chain.

*Example 4.4.* Assume that we want to use a Gibbs sampler to estimate the probability $\mathbb{P}(X_1 \geq 0, X_2 \geq 0)$ for a $\mathsf{N}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right)$ distribution.[5] The marginal distributions are

$$X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2) \qquad \text{and} \qquad X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$$

In order to construct a Gibbs sampler, we need the conditional distributions $X_1|X_2 = x_2$ and $X_2|X_1 = x_1$. We have[6]

$$\begin{aligned} f(x_1, x_2) \quad &\propto \quad \exp\left(-\frac{1}{2}\left(\begin{pmatrix}x_1\\ x_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)' \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}^{-1} \left(\begin{pmatrix}x_1\\ x_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)\right) \\ &\propto \quad \exp\left(-\frac{(x_1 - (\mu_1 + \sigma_{12}/\sigma_{22}^2(x_2 - \mu_2)))^2}{2(\sigma_1^2 - (\sigma_{12})^2/\sigma_2^2)}\right), \end{aligned}$$

i.e.

$$X_1|X_2 = x_2 \sim \mathsf{N}(\mu_1 + \sigma_{12}/\sigma_2^2(x_2 - \mu_2), \sigma_1^2 - (\sigma_{12})^2/\sigma_2^2)$$

Thus the Gibbs sampler for this problem consists of iterating for $t = 1, 2, \ldots$

1. Draw $X_1^{(t)} \sim \mathsf{N}(\mu_1 + \sigma_{12}/\sigma_2^2(X_2^{(t-1)} - \mu_2), \sigma_1^2 - (\sigma_{12})^2/\sigma_2^2)$

[5] A Gibbs sampler is of course not the optimal way to sample from a $\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. A more efficient way is: draw $Z_1, \ldots, Z_p \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and set $(X_1, \ldots, X_p)' = \boldsymbol{\Sigma}^{1/2}(Z_1, \ldots, Z_p)' + \boldsymbol{\mu}$

[6] We make use of

$$\begin{aligned} &\left(\begin{pmatrix}x_1\\ x_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)' \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}^{-1} \left(\begin{pmatrix}x_1\\ x_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) \\ = \ &\frac{1}{\sigma_1^2\sigma_2^2 - (\sigma_{12})^2}\left(\begin{pmatrix}x_1\\ x_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)' \begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix} \left(\begin{pmatrix}x_1\\ x_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) \\ = \ &\frac{1}{\sigma_1^2\sigma_2^2 - (\sigma_{12})^2}\left(\sigma_2^2(x_1 - \mu_1)^2 - 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2)\right) + \text{ const} \\ = \ &\frac{1}{\sigma_1^2\sigma_2^2 - (\sigma_{12})^2}\left(\sigma_2^2 x_1^2 - 2\sigma_2^2 x_1\mu_1 - 2\sigma_{12}x_1(x_2 - \mu_2)\right) + \text{ const} \\ = \ &\frac{1}{\sigma_1^2 - (\sigma_{12})^2/\sigma_2^2}\left(x_1^2 - 2x_1(\mu_1 + \sigma_{12}/\sigma_2^2(x_2 - \mu_2))\right) + \text{ const} \\ = \ &\frac{1}{\sigma_1^2 - (\sigma_{12})^2/\sigma_2^2}\left(x_1 - (\mu_1 + \sigma_{12}/\sigma_2^2(x_2 - \mu_2)\right)^2 + \text{ const} \end{aligned}$$



2. Draw $X_2^{(t)} \sim \mathsf{N}(\mu_2 + \sigma_{12}/\sigma_1^2(X_1^{(t)} - \mu_1), \sigma_2^2 - (\sigma_{12})^2/\sigma_1^2)$.

Now consider the special case $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_{12} = 0.3$. Figure 4.4 shows the sample paths of this Gibbs sampler.

Using theorem 4.6 we can estimate $\mathbb{P}(X_1 \geq 0, X_2 \geq 0)$ by the proportion of samples $(X_1^{(t)}, X_2^{(t)})$ with $X_1^{(t)} \geq 0$ and $X_2^{(t)} \geq 0$. Figure 4.3 shows this estimate. ◁

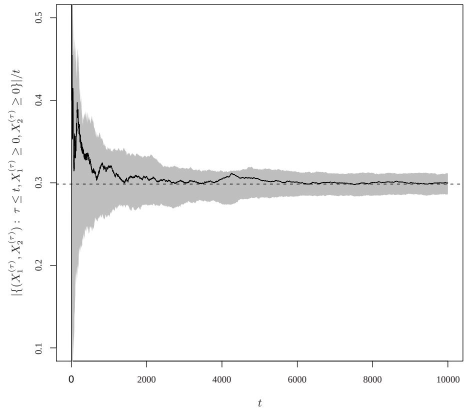
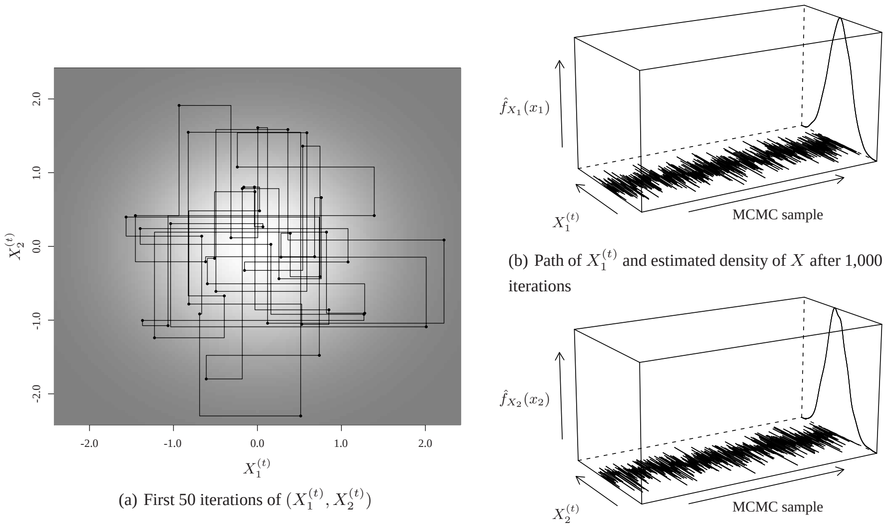



**Figure 4.3.** Estimate of the $\mathbb{P}(X_1 \geq 0, X_2 \geq 0)$ obtained using a Gibbs sampler. The area shaded in grey corresponds to the range of 100 replications.

Note that the realisations $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots)$ form a Markov chain, and are thus *not* independent, but typically positively correlated. The correlation between the $\mathbf{X}^{(t)}$ is larger if the Markov chain moves only slowly (the chain is then said to be *slowly mixing*). For the Gibbs sampler this is typically the case if the variables $X_j$ are strongly (positively or negatively) correlated, as the following example shows.

*Example 4.5 (Sampling from a highly correlated bivariate Gaussian).* Figure 4.5 shows the results obtained when sampling from a bivariate Normal distribution as in example 4.4, however with $\sigma_{12} = 0.99$. This yields a correlation of $\rho(X_1, X_2) = 0.99$. This Gibbs sampler is a lot slower mixing than the one considered in example 4.4 (and displayed in figure 4.4): due to the strong correlation the Gibbs sampler can only perform very small movements. This makes subsequent samples $X_j^{(t-1)}$ and $X_j^{(t)}$ highly correlated and thus yields to a slower convergence, as the plot of the estimated densities show (panels (b) and (c) of figures 4.4 and 4.5). ◁

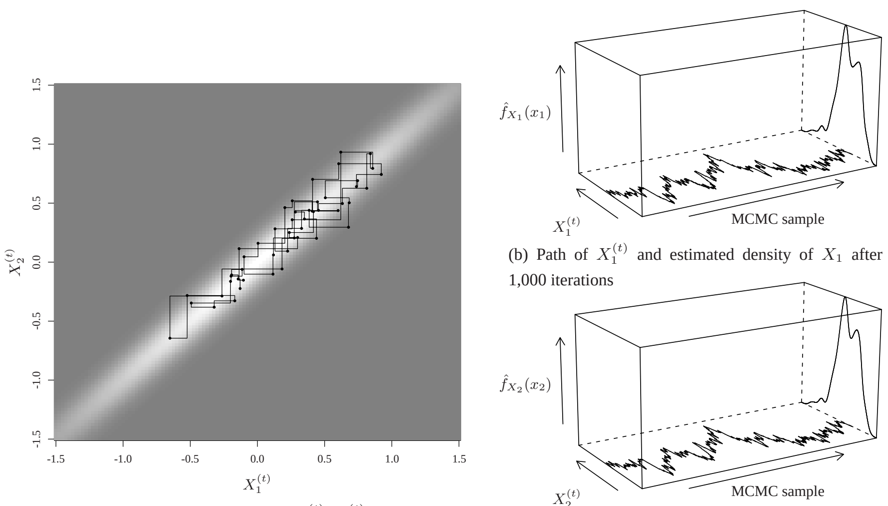



(a) First 50 iterations of $(X_1^{(t)}, X_2^{(t)})$

(b) Path of $X_1^{(t)}$ and estimated density of $X$ after 1,000 iterations

(c) Path of $X_2^{(t)}$ and estimated density of $X_2$ after 1,000 iterations

**Figure 4.4.** Gibbs sampler for a bivariate standard normal distribution with correlation $\rho(X_1, X_2) = 0.3$.




(a) First 50 iterations of $(X_1^{(t)}, X_2^{(t)})$

(b) Path of $X_1^{(t)}$ and estimated density of $X_1$ after 1,000 iterations

(c) Path of $X_2^{(t)}$ and estimated density of $X_2$ after 1,000 iterations

**Figure 4.5.** Gibbs sampler for a bivariate normal distribution with correlation $\rho(X_1, X_2) = 0.99$.