Chapter 13: Markov Chain Monte Carlo Methods–Part 1[a]

Recall that in Bayesian statistics all the inference is based on the posterior distribution. However, except for some very particular Bayesian models, the posterior distribution is intractable and therefore applying Bayesian methods in practice requires the existence of tools to approximate it.

Markov chain Monte Carlo (MCMC) algorithms are the most popular tools to approximate the posterior distributions arising in Bayesian statistics.

Given a distribution μ of interest, an MCMC algorithm generates a trajectory $\{z_t\}_{t=1}^T$ of a Markov chain $(Z_t)_{t\geq 1}$ having μ as invariant distribution, so that μ can be approximated by the empirical distribution $\frac{1}{T}\sum_{t=1}^T \delta_{z_t}$. The existence of MCMC methods dates back to Metropolis et al. (1953) [11] and Hasting (1970) [5]. However, it is only the (relatively) recent increase of the computational power that has made these sampling techniques (and hence Bayesian inference) widely applicable.

In this chapter we focus on three popular MCMC algorithms, namely the Metropolis-Hastings algorithm, the Gibbs sampler and the Metropolis-within-Gibbs algorithm.

We first present the Metropolis-Hastings algorithm, and develop its theory, in the case where the target distribution has a finite support. We then give its general version and state (without any proofs) the corresponding main theoretical result, before introducing the Gibbs sampler and the Metropolis-within-Gibbs algorithm.

[a]This chapter is essentially taken from the lecture notes Bayesian Modelling unit that I taught during 2016 and 2021, and is based on [16].

Markov chain on a finite state space: Notation

Let $\mathcal{Z} = \{1, \ldots, m\}$ for some $m \in \mathbb{N}$. Each $i \in \mathcal{Z}$ is called a state and $\mathcal{Z}$ is called the state space. Let $\mathcal{P}(\mathcal{Z})$ be the set of probability distributions on $\mathcal{Z}$.

We say that a matrix $\boldsymbol{P} = [p_{ij}]_{i,j=1}^m$ is stochastic if every row is a distribution on $\mathcal{Z}$; that is

$$\sum_{j=1}^m p_{ij} = 1, \quad \min_{j \in \mathcal{Z}} p_{ij} \geq 0, \quad \forall i \in \mathcal{Z}.$$

Definition 13.7 *We say that $(Z_t)_{t \geq 0}$ is a Markov chain with initial distribution $\lambda_0 \in \mathcal{P}(\mathcal{Z})$ and transition matrix $\boldsymbol{P}$ if*

1. *Z_0 has distribution λ_0*
2. *For all $t \geq 0$ and $(i_0, \ldots, i_{t+1}) \in \mathcal{Z}^{t+2}$,*

$$\mathbb{P}(Z_{t+1} = i_{t+1} | Z_t = i_t, \ldots, Z_0 = i_0) = \mathbb{P}(Z_{t+1} = i_{t+1} | Z_t = i_t)$$

3. *For all $t \geq 0$ and $(i, j) \in \mathcal{Z}^2$,*

$$\mathbb{P}(Z_{t+1} = j | Z_t = i) = p_{ij}.$$

We say that $(Z_t)_{t \geq 0}$ is Markov$(\lambda_0, \boldsymbol{P})$ in short.

For a Markov$(\lambda_0, \boldsymbol{P})$ process $(Z_t)_{t \geq 0}$ we use below the shorthand

$$p_{ij}(t) = \mathbb{P}(Z_t = j | Z_0 = i), \quad t \geq 1, \quad (i, j) \in \mathcal{Z}^2$$

and, for $t \geq 0$, we denote by λ_t the marginal distribution of Z_t; that is

$$\lambda_t := \big(\mathbb{P}(Z_t = 1), \ldots, \mathbb{P}(Z_t = m)\big) \in \mathcal{P}(\mathcal{Z}). \tag{13.1}$$

Lastly, for $t \geq 0$ we let $p_{ij}^{(t)}$ be the element (i, j) of $\boldsymbol{P}^t$, so that $\boldsymbol{P}^t = [p_{ij}^{(t)}]_{i,j=1}^m$.

Some key definitions and properties of a Markov$(\lambda_0, \boldsymbol{P})$ process

The following proposition collects two simple results.

Proposition 13.1 *For any $t \geq 1$, $p_{ij}(t) = p_{ij}^{(t)}$ for all $(i,j) \in \mathcal{Z}^2$ and, for all $t \geq 0$, $\lambda_t^\top = \lambda_0^\top \boldsymbol{P}^t$.*

Proof: We first show the first part of the proposition by induction. The result is trivially true for $t = 1$. Assume now that it holds for $t = t'$ for some $t' \geq 1$, and let $i, j \in \mathcal{Z}$. Then,

$$p_{ij}(t'+1) = \mathbb{P}(Z_{t'+1} = j | Z_0 = i) = \sum_{k=1}^{m} \mathbb{P}(Z_{t'+1} = j, Z_{t'} = k | Z_0 = i) = \sum_{k=1}^{m} \mathbb{P}(Z_{t'+1} = j | Z_{t'} = k) \mathbb{P}(Z_{t'} = k | Z_0 = i)$$
$$= \sum_{k=1}^{m} p_{kj} p_{ik}^{(t')} = [\boldsymbol{P}^{t'} \boldsymbol{P}]_{ij} = p_{ij}^{(t'+1)}.$$

We now show the second part of the proposition by induction. The result is trivially true for $t = 0$. Assume now that it holds for $t = t'$ for some $t' \geq 0$. Then,

$$\lambda_{t'+1,i} = \mathbb{P}(Z_{t'+1} = i) = \sum_{k=1}^{m} \mathbb{P}(Z_{t'+1} = i | Z_{t'} = k) \mathbb{P}(Z_{t'} = k) = \sum_{k=1}^{m} \lambda_{t',k} p_{ki}, \quad \forall i \in \mathcal{Z}$$

showing that $\lambda_{t'+1}^\top = \lambda_{t'}^\top \boldsymbol{P} = (\lambda_0^\top \boldsymbol{P}^{t'}) \boldsymbol{P} = \lambda_0^\top \boldsymbol{P}^{t'+1}$. □

Definition 13.8 *$\boldsymbol{P}$ is irreducible if for any $(i,j) \in \mathcal{Z}^2$ there exists a $t \geq 1$ such that $p_{ij}^{(t)} > 0$.*

In words, $\boldsymbol{P}$ is irreducible if for all $\lambda_0 \in \mathcal{P}(\mathcal{Z})$ a Markov$(\lambda_0, \boldsymbol{P})$ process $(Z_t)_{t \geq 0}$ can go to any state j from any state i.

Definition 13.9 *$\boldsymbol{P}$ is aperiodic if, for all $i \in \mathcal{Z}$, we have $p_{ii}^{(t)} > 0$ for all sufficiently large t.*

In words, $\boldsymbol{P}$ is aperiodic if for all $\lambda_0 \in \mathcal{P}(\mathcal{Z})$ a Markov$(\lambda_0, \boldsymbol{P})$ process $(Z_t)_{t \geq 0}$ can, for all $i \in \mathcal{Z}$, return to state i at irregular times when t is large enough.

Definition 13.10 *A probability distribution $\mu \in \mathcal{P}(\mathcal{Z})$ is invariant for $\boldsymbol{P}$ if $\mu^\top \boldsymbol{P} = \mu^\top$.*

Some key definitions and properties of Markov$(\lambda_0, \boldsymbol{P})$ process (end)

Remark: By Proposition 13.1 (second part), if μ is invariant for $\boldsymbol{P}$ and $(Z_t)_{t\geq 0}$ is Markov$(\mu, \boldsymbol{P})$ then $Z_t \sim \mu$ for all $t \geq 0$.

Proposition 13.2 *$\boldsymbol{P}$ has at least one invariant distribution.*

Proof: Recall that Brouwer's fixed-point theorem states that if f is a continuous mapping from a compact set $\mathcal{S}$ into itself then there exists an $x_0 \in \mathcal{K}$ such that $x_0 = f(x_0)$. Let $S = \mathcal{P}(\mathcal{Z})$ (i.e. $S = \{\mu \in [0,\infty)^m$ such that $\sum_{i=1}^m \mu_i = 1\}$) and f be the mapping $\mu \mapsto \boldsymbol{P}^\top \mu$. Note that S is closed and bounded (and therefore compact) while f is continuous on S. To apply Brouwer's fixed point theorem it remains to show that $f(S) = S$. For every $\mu \in S$ all the components of $f(\mu)$ are non-negative because $\min_{i,j\in\mathcal{Z}} p_{ij} \geq 0$ and $\min_{i\in\mathcal{Z}} \mu_i \geq 0$. Let $f_i(\mu)$ be the ith component of $f(\mu)$ and note that

$$\sum_{i=1}^m f_i(\mu) = 1_m^\top(\boldsymbol{P}^\top \mu) = (\boldsymbol{P}1_m)^\top \mu = 1_m^\top \mu = 1,$$

which concludes to show that $f(S) = S$. Therefore, by Brouwer's fixed point theorem, there exists a $\mu \in S$ such that $\mu = f(\mu) = \boldsymbol{P}^\top \mu \Leftrightarrow \mu^\top = \mu^\top \boldsymbol{P}$. The proof is complete. □

Remark: $\boldsymbol{P}$ can have more than one invariant distributions.

An invariant distribution μ of $\boldsymbol{P}$ is often called stationary/equilibrium distribution of $\boldsymbol{P}$ because of the following result.

Theorem 13.1 *Assume that, for at least one $i \in \mathcal{Z}$, $\lim_{t\to\infty} p_{ij}^{(t)}$ exists for all $j \in \mathcal{Z}$. Then,*

$$\mu := \Big(\lim_{t\to\infty} p_{ij}^{(t)},\ j \in \mathcal{Z} \Big)$$

is an invariant distribution of $\boldsymbol{P}$.

Proof: Since $p_{ij}^{(t)} \in [0,1]$ for all $t \geq 1$ and $j \in \mathcal{Z}$ then $\mu_j \in [0,1]$ for all $j \in \mathcal{Z}$. In addition[a],

$$\sum_{j=1}^m \mu_j = \sum_{j=1}^m \lim_{t\to\infty} p_{ij}^{(t)} = \lim_{t\to\infty} \sum_{j=1}^m p_{ij}^{(t)} = \lim_{t\to\infty} = 1$$

and thus $\mu \in \mathcal{P}(\mathcal{Z})$. To show that μ is an invariant distribution of $\boldsymbol{P}$ let $j \in \mathcal{Z}$ and note that

$$\sum_{k=1}^m \mu_k p_{kj} = \sum_{k=1}^m \lim_{t\to\infty} p_{ik}^{(t)} p_{kj} = \lim_{t\to\infty} \sum_{k=1}^m p_{ik}^{(t)} p_{kj} = \lim_{t\to\infty} p_{ij}^{(t+1)} = \mu_j.$$

□

[a] Recall that we can swap a limit and a sum when the sum is over a finite number of terms.

Convergence of Markov chains to equilibrium

Theorem 13.2 *Let $\boldsymbol{P}$ be an irreducible and aperiodic stochastic matrix with invariant distribution $\mu \in \mathcal{P}(\mathcal{Z})$. Let $\lambda_0 \in \mathcal{P}(\mathcal{Z})$ and $(Z_t)_{t\geq 0}$ be a Markov$(\lambda_0, \boldsymbol{P})$ process. Then, there exist constants $\rho \in (0,1)$ and $c \in (0,\infty)$ such that, for all $(i,j) \in \mathcal{Z}^2$ and $t \geq 0$,*

$$|p_{ij}^{(t)} - \mu_j| \leq c\,\rho^t \quad \text{and} \quad \big|\mathbb{P}(Z_t = j) - \mu_j\big| \leq c\,\rho^t.$$

Remark: By Proposition 13.2 every stochastic matrix $\boldsymbol{P}$ has at least one invariant distribution. Thus, Theorem 13.2 implies that if $\boldsymbol{P}$ is irreducible and aperiodic then $\boldsymbol{P}$ has a unique invariant distribution.

Remark: Theorem 13.2 implies that if $(Z_t)_{t\geq 0}$ is Markov$(\lambda_0, \boldsymbol{P})$ for an irreducible and aperiodic stochastic matrix $\boldsymbol{P}$ then, as $t \to \infty$, $Z_t \xrightarrow{\text{D}} \mu$ where μ is the unique invariant distribution of $\boldsymbol{P}$.

To illustrate the conclusion of Theorem 13.2, let $(\epsilon, \delta) \in [0,1]^2$, $m = 2$ and $\boldsymbol{P} = \begin{pmatrix} \epsilon & 1-\epsilon \\ 1-\delta & \delta \end{pmatrix}$. Remark that the matrix $\boldsymbol{P}$ is irreducible and aperiodic for all $(\epsilon, \delta) \in (0,1)^2$.

Figure 13.1 below shows the behaviour of $p_{11}^{(t)}$ as t increases for $(\epsilon, \delta) = (0.2, 0.375)$ and for $(\epsilon, \delta) = (10^{-4}, 0.375)$.

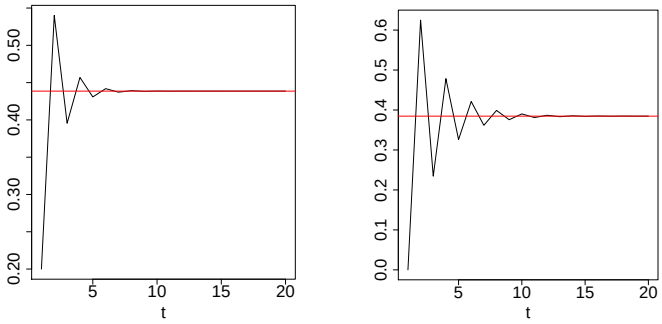

Figure 13.1: Convergence of $p_{11}^{(t)}$ towards μ_1 (horizontal line) for $(\epsilon, \delta) = (0.2, 0.375)$ (left plot) and for $(\epsilon, \delta) = (10^{-4}, 0.375)$.

Proof of Theorem 13.2[a]

We start with a preliminary result.

Lemma 13.1 *Let $\boldsymbol{P}$ be irreducible and aperiodic. Then, there exists a $t_0 \geq 1$ such that $\min_{(i,j)\in\mathcal{Z}^2} p_{ij}^{(t)} > 0$ for all $t \geq t_0$.*

Proof: Let $(i,j,k) \in \mathcal{Z}^3$ and remark first that, as $\boldsymbol{P}$ is irreducible, there exist a $t_1 \geq 1$ and a $t_2 \geq 1$ such that $p_{ij}^{(t_1)} > 0$ and $p_{ki}^{(t_2)} > 0$. In addition, as $\boldsymbol{P}$ is aperiodic, there exists a $t_3 \geq 1$ such that $p_{ii}^{(t)} > 0$ for all $t \geq t_3$.

Let $(Z_t)_{t\geq 0}$ be Markov$(\lambda_0, \boldsymbol{P})$. Then, using the above observations, the Markov property of $(Z_t)_{t\geq 0}$ and Proposition 13.1, for any $t \geq t_3$ we have

$$\begin{aligned}
p_{kj}^{(t_1+t_2+t)} &= \mathbb{P}(Z_{t_1+t_2+t} = j | Z_0 = k) \\
&\geq \mathbb{P}(Z_{t_1+t_2+t} = j, Z_{t_2+t} = i, Z_{t_2} = i | Z_0 = k) \\
&= \mathbb{P}(Z_{t_1+t_2+t} = j | Z_{t_2+t} = i)\mathbb{P}(Z_{t_2+t} = i | Z_{t_2} = i)\mathbb{P}(Z_{t_2} = i | Z_0 = k) \\
&= p_{ij}^{(t_1)}\, p_{ii}^{(t)}\, p_{ki}^{(t_2)} \\
&> 0
\end{aligned}$$

showing that $p_{kj}^{(t)} > 0$ for all $t \geq t_1 + t_2 + t_3$. The result follows. □

Proof of Theorem 13.2: We first assume that $\min_{(i,j)\in\mathcal{Z}^2} p_{ij} > 0$.

Let $\boldsymbol{\Pi} = 1_m \mu^\top$ and note that

$$\boldsymbol{P}\boldsymbol{\Pi} = \boldsymbol{\Pi}\boldsymbol{P} = \boldsymbol{\Pi}^2 = \boldsymbol{\Pi}. \tag{13.2}$$

Note also that, because $\min_{(i,j)\in\mathcal{Z}^2} p_{ij} > 0$, there exists an $\alpha \in (0,1)$ such that all the elements of the matrix $\boldsymbol{P} - \alpha\boldsymbol{\Pi}$ are non-negative. Let

$$\tilde{\boldsymbol{P}} = \frac{1}{1-\alpha}(\boldsymbol{P} - \alpha\boldsymbol{\Pi})$$

so that, since all the elements of $\tilde{\boldsymbol{P}}$ are non-negative and $\tilde{\boldsymbol{P}} 1_m = 1_m$, $\tilde{\boldsymbol{P}}$ is a stochastic matrix. Note also that, by (13.2),

$$\tilde{\boldsymbol{P}}\boldsymbol{\Pi} = \boldsymbol{\Pi}\tilde{\boldsymbol{P}} = \boldsymbol{\Pi}. \tag{13.3}$$

[a] This proof is due to Prof. Balint Toth.

Proof of Theorem 13.2 (end)

Next, let $t \geq 1$ and note that

$$\begin{aligned} \boldsymbol{P}^t = \Big((1-\alpha)\tilde{\boldsymbol{P}} + \alpha\boldsymbol{\Pi}\Big)^t &= (1-\alpha)^t\tilde{\boldsymbol{P}}^t + \sum_{s=1}^{t}\binom{t}{s}(1-\alpha)^{t-s}\tilde{\boldsymbol{P}}^{t-s}\alpha^s\Pi^s \\ &= (1-\alpha)^t\tilde{\boldsymbol{P}}^t + \boldsymbol{\Pi}\sum_{s=1}^{t}\binom{t}{s}(1-\alpha)^{t-s}\alpha^s \\ &= (1-\alpha)^t\tilde{\boldsymbol{P}}^t + \boldsymbol{\Pi}\big(1-(1-\alpha)^t\big) \end{aligned} \tag{13.4}$$

where the second equality uses the Binomial expansion (that holds because the matrices $\tilde{\boldsymbol{P}}$ and $\boldsymbol{\Pi}$ commute, by (13.3)), the third equality uses (13.2)-(13.3) and the last equality uses the fact that the sum is equal to $1 - \mathbb{P}(Z = 0)$ where $Z \sim \text{Binomial}(t, \alpha)$.

To proceed further for a matrix $\boldsymbol{A} = [a_{ij}]_{i,j}$ we let $|\boldsymbol{A}| = [|a_{ij}|]_{i,j}$. Then, using (13.4),

$$|\boldsymbol{P}^t - \boldsymbol{\Pi}| = (1-\alpha)^t|\tilde{\boldsymbol{P}}^t - \boldsymbol{\Pi}| \leq (1-\alpha)^t 1_m 1_m^T, \quad \forall t \geq 1 \tag{13.5}$$

where the inequality holds because $\tilde{\boldsymbol{P}}$ and $\boldsymbol{\Pi}$ are both stochastic matrices. Finally, since $\boldsymbol{P}^t - \boldsymbol{\Pi} = [p_{ij}^{(t)} - \mu_j]_{i,j=1}^m$, inequality (13.5) yields

$$|p_{ij}^{(t)} - \mu_j| \leq (1-\alpha)^t, \quad \forall (i,j) \in \mathcal{Z}^2, \quad \forall t \geq 1$$

showing that the conclusion of Theorem 13.2 holds with $c = 1$ and $\rho = (1-\alpha)$ in the special case where $\min_{(i,j)\in\mathcal{Z}^2} p_{ij} > 0$.

We now prove the theorem assuming that $\min_{(i,j)\in\mathcal{Z}^2} p_{ij} = 0$.

Since $\boldsymbol{P}$ is irreducible and aperiodic, there exists by Lemma 13.1, a $t_0 \geq 1$ such that $\min_{(i,j)\in\mathcal{Z}^2} p_{ij}^{(t)} > 0$ for all $t \geq t_0$. Let $\boldsymbol{P}_0 = \boldsymbol{P}^{t_0}$ so that, as per above, there exists an $\alpha_0 \in (0,1)$ such that all the elements of the matrix $\boldsymbol{P}_0 - \alpha_0\boldsymbol{\Pi}$ are non-negative.

Repeating the above computations with $\boldsymbol{P}$ replaced by $\boldsymbol{P}_0$ and α replaced by α_0, we have, noting that $\boldsymbol{P}_0^t$ has elements $[p_{ij}^{(t_0+t)}]_{i,j=1}^M$,

$$|p_{ij}^{(t_0+t)} - \mu_j| \leq (1-\alpha_0)^t, \quad \forall (i,j) \in \mathcal{Z}^2, \quad \forall t \geq 1.$$

or, equivalently,

$$|p_{ij}^{(t)} - \mu_j| \leq (1-\alpha_0)^{t-t_0}, \quad \forall (i,j) \in \mathcal{Z}^2, \quad \forall t > t_0.$$

Then, as $|p_{ij}^{(t)} - \mu_j| \leq 1$ for all $(i,j) \in \mathcal{Z}^2$ and all $t \geq 0$, the conclusion of Theorem 13.2 holds with $c = (1-\alpha_0)^{-t_0}$ and $\rho = (1-\alpha_0)$. □

Estimation of expectations under μ

Let $\boldsymbol{P}$ and μ be as in Theorem 13.2. Then, in practice we are often interested in estimating $\mu(\varphi) := \sum_{i=1}^{m} \varphi(i)\mu_i$ for some $\varphi : \mathcal{Z} \to \mathbb{R}$.

The following result shows that, under the assumptions of Theorem 13.2, if $(Z_t)_{t\geq 0}$ is a Markov$(\lambda_0, \boldsymbol{P})$ process then the estimator

$$\frac{1}{T}\sum_{t=1}^{T} \varphi(Z_t)$$

converges to $\mu(\varphi)$ at the standard $T^{-1/2}$ rate as $T \to \infty$.

Corollary 13.1 *Consider the set-up of Theorem 13.2 and let $\varphi : \mathcal{Z} \to \mathbb{R}$. Then, there exists a constant $C \in (0, \infty)$ such that*

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T} \varphi(Z_t) - \sum_{i=1}^{m} \varphi(i)\mu_i\Big)^2\Big]^{1/2} \leq \frac{C}{T^{1/2}}, \quad \forall T \geq 1.$$

Remark: It can be shown that a strong law of large numbers holds, namely that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \varphi(Z_t) \to \sum_{i=1}^{m} \varphi(i)\mu_i, \quad \mathbb{P}\text{-almost surely.}$$

Proof of Corollary 13.1

We first show the following result

Lemma 13.2 *Consider the set-up of Theorem 13.2. Then, there exists a constant $C_1 \in (0,\infty)$ such that*

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{i\}}(Z_t)-\mu_i\Big)^2\Big] \leq \frac{C_1}{T}, \quad \forall i \in \mathcal{Z}, \quad \forall T \geq 1.$$

Proof: Fix i. Then, $\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{i\}}(Z_t)-\mu_i\Big)^2\Big] = \frac{1}{T}\big(v_1(T)+v_2(T)\big)$ with $v_1(T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)^2\big]$ and with

$$v_2(T) = \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)(\mathbf{1}_{\{i\}}(Z_s)-\mu_i)\big].$$

Then, to prove Lemma 13.2 it suffices to show that $\sup_{T\geq 1} v_1(T) < \infty$ and that $\sup_{T\geq 1}|v_2(T)| < \infty$.

For $v_1(T)$ we trivially have $v_1(T) \leq 1$ and thus, as required, $\sup_{T\geq 1} v_1(T) < \infty$.

We now study $v_2(T)$. To this aim remark first that, by Proposition 13.1 and using the law of iterated expectations, for $1 \leq t < s$,

$$\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)(\mathbf{1}_{\{i\}}(Z_s)-\mu_i)\big] = \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)\mathbb{E}[(\mathbf{1}_{\{i\}}(Z_s)-\mu_i)|Z_t]\big] = \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)(p_{Z_t i}^{(s-t)}-\mu_i)\big].$$

Therefore,

$$\begin{aligned}|v_2(T)| &\leq \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\Big|\mathbb{E}\Big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)(\mathbf{1}_{\{i\}}(Z_s)-\mu_i)\Big]\Big| = \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\Big|\mathbb{E}\Big[(\mathbf{1}_{\{i\}}(Z_t)-\mu_i)(p_{Z_t i}^{(s-t)}-\mu_i)\Big]\Big| \\ &\leq \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}\Big[|\mathbf{1}_{\{i\}}(Z_t)-\mu_i|\,|p_{Z_t i}^{(s-t)}-\mu_i|\Big] \leq \frac{2\,c}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\rho^{s-t} \leq \frac{2\,c}{T}\sum_{t=1}^{T-1}\sum_{s=1}^{\infty}\rho^{s} \leq \frac{2\,c}{1-\rho}\end{aligned}$$

where the first inequality uses the triangle inequality, the second inequality uses Jensen's inequality and the third inequality the result of Theorem 13.2. The proof of Lemma 13.2 is complete. □

Proof of Corollary 13.1: $\varphi : \mathcal{Z} \to \mathbb{R}$ and $\mu(\varphi) = \sum_{i=1}^{m}\varphi(i)\mu_i$.

Then,

$$\frac{1}{T}\sum_{t=1}^{T}\varphi(Z_t) - \sum_{i=1}^{m}\varphi(i)\mu_i = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{m}\varphi(i)\big(\mathbf{1}_{\{i\}}(Z_t)-\mu_i\big) = \sum_{i=1}^{m}\varphi(i)\frac{1}{T}\sum_{t=1}^{T}\big(\mathbf{1}_{\{i\}}(Z_t)-\mu_i\big)$$

so that, using Cauchy-Schwartz's inequality,

$$\Big(\frac{1}{T}\sum_{t=1}^{T}\varphi(Z_t) - \sum_{i=1}^{m}\varphi(i)\mu_i\Big)^2 \leq \sum_{i=1}^{m}\varphi(i)^2\sum_{i=1}^{m}\Big(\frac{1}{T}\sum_{t=1}^{T}\big(\mathbf{1}_{\{i\}}(Z_t)-\mu_i\big)\Big)^2$$

and therefore

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\varphi(Z_t) - \sum_{i=1}^{m}\varphi(i)\mu_i\Big)^2\Big] \leq \sum_{i=1}^{m}\varphi(i)^2\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{i\}}(Z_t)-\mu_i\Big)^2\Big]$$

and the result follows from Lemma 13.2. □

Building a Markov chain having a given invariant distribution

Above $\boldsymbol{P}$ was given and we have shown that if $\boldsymbol{P}$ is irreducible and aperiodic then a Markov$(\lambda_0, \boldsymbol{P})$ process converges to the unique invariant distribution μ of $\boldsymbol{P}$.

We now consider the converse problem: Given μ, the distribution we are interested in, we now want to construct a transition matrix $\boldsymbol{P}$ such that a Markov$(\lambda_0, \boldsymbol{P})$ process converges to μ.

To solve this problem we start with the following simple lemma.

Lemma 13.3 *Let $\mu \in \mathcal{P}(\mathcal{Z})$ and assume that there exists a transition matrix $\boldsymbol{P}$ such that*

$$\mu_i p_{ij} = \mu_j p_{ji}, \quad \forall (i,j) \in \mathcal{Z}^2. \tag{13.6}$$

Then, μ is an invariant distribution of $\boldsymbol{P}$.

Proof: If (13.6) holds then $\sum_{i=1}^m \mu_i p_{ij} = \sum_{i=1}^m \mu_j p_{ji} = \mu_j \sum_{i=1}^m p_{ji} = \mu_j$ for all $j \in \mathcal{Z}$ and the result follows. □

Remark: Condition (13.6) is known as the detailed balance condition. It implies that, at equilibrium (i.e. when $Z_t \sim \mu$), the joint probability $\mathbb{P}(Z_t = i, Z_{t+1} = j)$ is symmetric in t and $t+1$.

Remark: When there exists a $\mu \in \mathcal{P}(\mathcal{Z})$ such that (13.6) holds we say that the Markov$(\lambda_0, \boldsymbol{P})$ process is reversible.

Lemma 13.3 shows that to construct a Markov chain having μ as invariant distribution it is enough to construct a transition matrix $\boldsymbol{P}$ such that (13.6) holds. Surprisingly, not only it is always feasible to construct such a matrix $\boldsymbol{P}$, but it is (very) easy using the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm: Main theorem

Theorem 13.3 *Let $\boldsymbol{Q} = [q_{ij}]_{i,j=1}^m$ be a transition matrix such that $q_{ij} > 0$ for all $(i,j) \in \mathcal{Z}^2$, $\mu \in \mathcal{P}(\mathcal{Z})$ be such that $\mu_i > 0$ for all $i \in \mathcal{Z}$ and $\boldsymbol{P}^{\mathrm{MH}} = [p_{ij}^{\mathrm{MH}}]_{i,j=1}^m$ be such that, for every $i \in \mathcal{Z}$,*

$$p_{ij}^{\mathrm{MH}} = \begin{cases} q_{ij} \, \min\left\{1, \frac{\mu_j \, q_{ji}}{\mu_i \, q_{ij}}\right\}, & j \neq i \\ 1 - \sum_{k \neq i} q_{ik} \, \min\left\{1, \frac{\mu_k \, q_{ki}}{\mu_i \, q_{ik}}\right\}, & j = i. \end{cases}$$

Then, the transition matrix $\boldsymbol{P}^{\mathrm{MH}}$ is irreducible, aperiodic and has μ as unique invariant distribution.

Proof: For all $i \neq j$ we have

$$\begin{aligned} \mu_j p_{ij}^{\mathrm{MH}} = \mu_i q_{ij} \, \min\left\{1, \frac{\mu_j \, q_{ji}}{\mu_i \, q_{ij}}\right\} &= \min\left\{\mu_i q_{ij}, \mu_j \, q_{ji}\right\} \\ &= \mu_j q_{ji} \, \min\left\{1, \frac{\mu_i \, q_{ij}}{\mu_j \, q_{ji}}\right\} \\ &= \mu_i p_{ji}^{\mathrm{MH}} \end{aligned} \tag{13.7}$$

and thus, by Lemma 13.3, $\boldsymbol{P}^{\mathrm{MH}}$ has μ as invariant distribution.

Under the assumptions on $\boldsymbol{Q}$ and μ, we have $p_{ij}^{\mathrm{MH}} \in (0,1)$ for all $i, j \in \mathcal{Z}$, and thus $\boldsymbol{P}^{\mathrm{MH}}$ is irreducible an aperiodic.

Finally, since $\boldsymbol{P}^{\mathrm{MH}}$ is irreducible an aperiodic, it follows by Theorem 13.2 that μ is the unique invariant distribution of $\boldsymbol{P}^{\mathrm{MH}}$. □

The Metropolis-Hastings algorithm: The algorithm

Metropolis-Hastings algorithm on a finite state space (Algorithm A1)

Input: $\mu \in \mathcal{P}(\mathcal{Z})$, $z_0 \in \mathcal{Z}$ and a transition matrix $\boldsymbol{Q}$ on $\mathcal{Z}$

Set $Z_0 = z_0$

for $t \geq 1$ **do**

$\tilde{Z}_t \sim Q(Z_{t-1}, \mathrm{d}\tilde{z}_t)$

Set $Z_t = \tilde{Z}_t$ with probability $\alpha(Z_{t-1}, \tilde{Z}_t)$ and $Z_t = Z_{t-1}$ with probability $1 - \alpha(Z_{t-1}, \tilde{Z}_t)$.

end for

Notation: For $z \in \mathcal{Z}$ the notation $\tilde{Z} \sim Q(z, \mathrm{d}\tilde{z})$ means that the random variable $\tilde{Z}$ is such that $\mathbb{P}(\tilde{Z} = j) = q_{zj}$, while the mapping $\alpha : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$ is defined by

$$\alpha(i, j) = \min\left\{1, \frac{\mu_j \, q_{ji}}{\mu_i \, q_{ij}}\right\}, \quad i, j \in \mathcal{Z}.$$

Remark: By Theorem 13.3, Algorithm A1 defines a Markov$(\delta_{z_0}, P_\mu^{\mathrm{MH}})$ process having μ as unique invariant distribution and such that the result of Theorem 13.2 and of Corollary 13.1 hold.

The Metropolis-Hastings algorithm on a general state space

The extension of the Metropolis-Hastings (M-H) algorithm on a finite state space to an arbitrary state space $\mathcal{Z}$ is straightforward:

Metropolis-Hastings algorithm
(Algorithm A2)

Input: $\mu \in \mathcal{P}(\mathcal{Z})$, $z_0 \in \mathcal{Z}$ and a Markov kernel Q on $\mathcal{Z}$.

Set $Z_0 = z_0$

for $t \geq 1$ **do**

$\tilde{Z}_t \sim Q(Z_{t-1}, \mathrm{d}\tilde{z}_t)$

Set $Z_t = \tilde{Z}_t$ with probability

$$\alpha(Z_{t-1}, \tilde{Z}_t) = \min\left\{1, \frac{\mu(\tilde{Z}_t) q(Z_{t-1}|\tilde{Z}_t)}{\mu(Z_{t-1}) q(\tilde{Z}_t|Z_{t-1})}\right\}$$

and $Z_t = Z_{t-1}$ otherwise.

end for

Remark: We abandon the notion of stochastic matrix to adopt the more general one of Markov kernel.

Notation: $q(\tilde{z}|z)$ is the density of the distribution $Q(z, \mathrm{d}\tilde{z})$, which is often called the proposal distribution.

Key remark: The M-H Algorithm A2 only requires to be able to compute $\mu(z)$ up to a normalizing constant.

Invariant distributions, irreducibility and aperiodicity for general state spaces

We start with the general definition of an invariant distribution.

Definition 13.11 *A probability distribution $\mu \in \mathcal{P}(\mathcal{Z})$ is an invariant distribution for the Markov kernel P if*

$$\int_{\mathcal{Z}} p(z'|z)\mu(z)\mathrm{d}y = \mu(z'), \quad \forall z' \in \mathcal{Z}.$$

Recall that, when $\mathcal{Z}$ is finite, a transition matrix $\boldsymbol{P}$ is irreducible if a Markov$(\lambda_0, \boldsymbol{P})$ process can go to any state $j \in \mathcal{Z}$ from any state $i \in \mathcal{Z}$. If $\mathcal{Z}$ is a continuous state space such a requirement is impossible to fulfil and we therefore need to weaken the notion of irreducibility.

Definition 13.12 *Given a probability distribution φ, the Markov chain $(Z_t)_{t\geq 0}$ with Markov kernel P is φ-irreducible if, for every $z \in \mathcal{Z}$ and set $A \subset \mathcal{Z}$ with $\varphi(A) > 0$, there exists a $t \geq 0$ such that $P^t(z, A) > 0$.*

In words, the Markov kernel P is φ-irreducible if it is possible to go to any set $A \subset \mathcal{Z}$ with $\varphi(A) > 0$ from any state $z \in \mathcal{Z}$.

Similarly, the notion of aperiodicity needs to be weakened.

Definition 13.13 *A Markov chain $(Z_t)_{t\geq 0}$ with Markov kernel P and stationary distribution μ is aperiodic if there do not exist a $p \geq 2$ and disjoints subsets $\mathcal{Z}_1, \ldots, \mathcal{Z}_p \subset \mathcal{Z}$ with $P(z, \mathcal{Z}_{i+1}) = 1$ for all $z \in \mathcal{Z}_i$ ($i \in \{1, \ldots, p-1\}$), $P(z, \mathcal{Z}_1) = 1$ for all $z \in \mathcal{Z}_p$, and such that $\mu(\mathcal{Z}_1) > 0$ (and hence $\mu(\mathcal{Z}_i) > 0$ for all i).*

A general convergence result for M-H algorithms

We first show that the M-H Algorithm A2 indeed defines a Markov chain having μ as invariant distribution.

Lemma 13.4 *Let $\mu \in \mathcal{P}(\mathcal{Z})$ and assume that there exists a Markov kernel P on $\mathcal{Z}$ such that*

$$\mu(z)p(\tilde{z}|z) = \mu(\tilde{z})p(z|\tilde{z}), \quad \forall (z,\tilde{z}) \in \mathcal{Z}^2. \tag{13.8}$$

Then, μ is an invariant distribution of P.

Proof: The result is proved by integrating both side of (13.8) with respect to z □

Corollary 13.2 *The Markov chain $(Z_t)_{t\geq 0}$ defined by the M-H Algorithm (A2) admits μ as invariant distribution.*

Proof: The result follows from Lemma 13.4, using similar computations as in (13.7) and noting that the Markov kernel P^{MH} of $(Z_t)_{t\geq 0}$ is defined by

$$P^{\mathrm{MH}}(z, \mathrm{d}\tilde{z}) = Q(z, \mathrm{d}\tilde{z})\alpha(z,\tilde{z}) + \delta_z(\mathrm{d}\tilde{z})\Big(1 - \int_{\mathcal{Z}} Q(z, \mathrm{d}\tilde{z})\alpha(z,\tilde{z})\Big).$$

□

The following result provides a simple way to check the validity of the M-H Algorithm A2 .

Theorem 13.4 ([16],Theorem 7.4, p.274) *Consider the M-H Algorithm A2. Assume that $Q(z, A) > 0$ for all $z \in \mathcal{Z}$ and all set $A \subset \mathcal{Z}$ such that $\mu(A) > 0$, and that*

$$\mathbb{P}\left(\frac{\mu(\tilde{Z}_t)q(Z_{t-1}|\tilde{Z}_t)}{\mu(Z_{t-1})q(\tilde{Z}_t|Z_{t-1})} < 1\right) > 0, \quad \forall t \geq 1.$$

Then, the resulting Markov chain $(Z_t)_{t\geq 0}$ is μ-irreducible and aperiodic, and consequently $\lim_{t\to\infty} \mathbb{P}(Z_t \in A) = \mu(A)$ for all (measurable) set $A \subset \mathcal{Z}$.

Central limit theorem for Markov chains

Let $(Z_t)_{t\geq 0}$ and $(Z'_t)_{t\geq 0}$ be two Markov chains defined by the M-H Algorithm A2 where the former is such that $Z_0 = z_0$ for some $z_0 \in \mathcal{Z}$ while the latter is such that $Z'_0 \sim \mu$ (the proposal distribution $Q(z, \mathrm{d}\tilde{z})$ being the same for the two processes).

Let $\mu(\varphi) := \int_{\mathcal{Z}} \varphi(z)\mu(z)\mathrm{d}z$ for some function $\varphi : \mathcal{Z} \to \mathbb{R}$ verifying $\mu(\varphi^2) < \infty$ and let

$$\hat{\mu}_T(\varphi) = \frac{1}{T}\sum_{t=1}^{T} \varphi(Z_t) \tag{13.9}$$

be an estimator of $\mu(\varphi)$.

Then, under some conditions, the following central limit theorem holds

$$\sqrt{T}\,\frac{\hat{\mu}_T(\varphi) - \mu(\varphi)}{\sigma} \xrightarrow{\mathrm{D}} \mathcal{N}_1(0,1), \quad \text{as } T \to \infty \tag{13.10}$$

with

$$\sigma^2 = \mathrm{Var}(\varphi(Z'_0)) + 2\sum_{t=1}^{\infty} \mathrm{Cov}\big(\varphi(Z'_0), \varphi(Z'_t)\big) = \mathrm{Var}(\varphi(Z'_0))\tau_\varphi$$

and where $\tau_\varphi = 1 + 2\sum_{t=1}^{\infty} \mathrm{Corr}\big(\varphi(Z'_0), \varphi(Z'_t)\big)$ is called the integrated auto-correlation time.

Remark: The asymptotic variance depends only on Q and not on z_0 (which is intuitive).

Choosing the proposal distribution $Q(z, \mathrm{d}\tilde{z})$

The choice of the proposal distribution $Q(z, \mathrm{d}\tilde{z})$ is important because

1. It influences the mixing time of the Markov chain, that is the time it needs to be 'close' to its stationary distribution (which is determined by the constant ρ in Theorem 13.2).
2. It influences the asymptotic variance of the estimator $\hat{\mu}_T(\varphi)$ of $\mu(\varphi)$ through the integrated auto-correlation time τ_φ.

The first point is particularly important because in practice we can only run Algorithm A2 for a finite number T of iterations. If Q is poorly chosen then the distribution of Z_T will be 'far' from the equilibrium distribution μ and the output of the algorithm will do a poor job at approximating μ.

Finding a good proposal distribution Q is both difficult and problem dependent: a given Q may perform well for some target distributions and very poorly for others.

In practice, the only solution is often to tune Q manually; that is, to try different proposal distributions until the output of the Algorithm (A2) suggests that the algorithm has converged (in the sense that Z_T is approximatively distributed according to μ).

Of course, we can never be sure that the M-H algorithm has converged but there are some ways to detect bad choices for Q, and notably the inspection of

1. the acceptance rate
2. the trace plots
3. the autocorrelation functions.

We explain these three complementary approaches in what follows.

Assessing the convergence using the acceptance rate

The first approach that can be used to assess the convergence of the M-H algorithm is to look at the acceptance rate:

$$r_T := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(z_t = \tilde{z}_t).$$

Indeed,

- A low value for this quantity indicates that the simulated trajectory $\{z_t\}_{t=1}^T$ remains for a long time at a given location before moving to a new state.
- A high acceptance rate usually (but not always) arises when Q is such that, with high probability, $\tilde{Z}_t$ is very 'close' to Z_{t-1}.

Hence, both a low and a large value of the acceptance rate is a sign that the mixing time of the Markov chain (i.e. the time needed to be close to equilibrium) is large and that the correlation between Z_t and Z_{t-k} is important even for large values of k.

There exist theoretical results suggesting that the "optimal" acceptance rate is 0.234 [17]. In practice, choosing a proposal distribution Q such that $r_T \approx 0.234$ works well.

The main advantage of this approach is its simplicity. However, by summarizing the behaviour of the Markov chain using a single number, the acceptance rate may hide important differences in the mixing times of the chain along its different coordinates.

Assessing the convergence using the trace plots and the autocorrelation functions

Let $Z_{i,t}$ be the ith coordinate of Z_t.

Then,

- The trace plot represents the simulated trajectory $\{z_{i,t}\}_{t=1}^{T}$ as a function of t.
- The auto-correlation function (ACF) returns, for an integer $k \geq 0$, an estimate $\hat{\gamma}_T(k)$ of $\text{Corr}(Z'_{i,0}, Z'_{i,k})$ where, as per above, the Markov chain $(Z'_t)_{t\geq 0}$ has the same Markov kernel as $(Z_t)_{t\geq 0}$ but is such that $Z'_0 \sim \mu$.

  For instance,

$$\hat{\gamma}_T(k) = \frac{\frac{1}{T-k}\sum_{s=k+1}^{T}\left(z_{i,t} - \bar{z}_{i,T}\right)\left(z_{i,t-k} - \bar{z}_{i,T}\right)}{\sqrt{\frac{1}{T}\sum_{s=1}^{T}\left(z_{i,t} - \bar{z}_{i,T}\right)^2 \frac{1}{T-k}\sum_{s=k+1}^{T}\left(z_{i,t-k} - \bar{z}_{i,T}\right)^2}}$$

  with

$$\bar{z}_{i,T} = \frac{1}{T}\sum_{t=1}^{T} z_{i,t}.$$

Looking at the trace plots and at the autocorrelation function therefore allows to assess the convergence of the Markov chain coordinate by coordinate.

The M-H algorithm: A toy example

Let $\mathcal{Z} = \mathbb{R}$, $\mu(z)\mathrm{d}z = \mathcal{N}_1(0,1)$ and $Q_\sigma(z, \mathrm{d}\tilde{z}) = \mathcal{N}_1(z, \sigma^2)$. For this example the M-H Algorithm A2 is used with $Q = Q_\sigma$ and with starting $z_0 = 3$.

The trace plots and ACFs obtained $\sigma = 0.01$, $\sigma = 5$ and $\sigma = 100$ are shown in Figure 13.2 below. The corresponding values for the acceptance rate r_T are $r_T = 0.98$, $r_T = 0.0114$ and $r_T = 0.24$.

From this figure we notably see when the Markov chain is highly autocorrelated the acceptance rate is either low or high. It is also worth noting that for $\sigma = 0.01$ the Markov chain has not reached the mode of μ after 5 000 iterations.

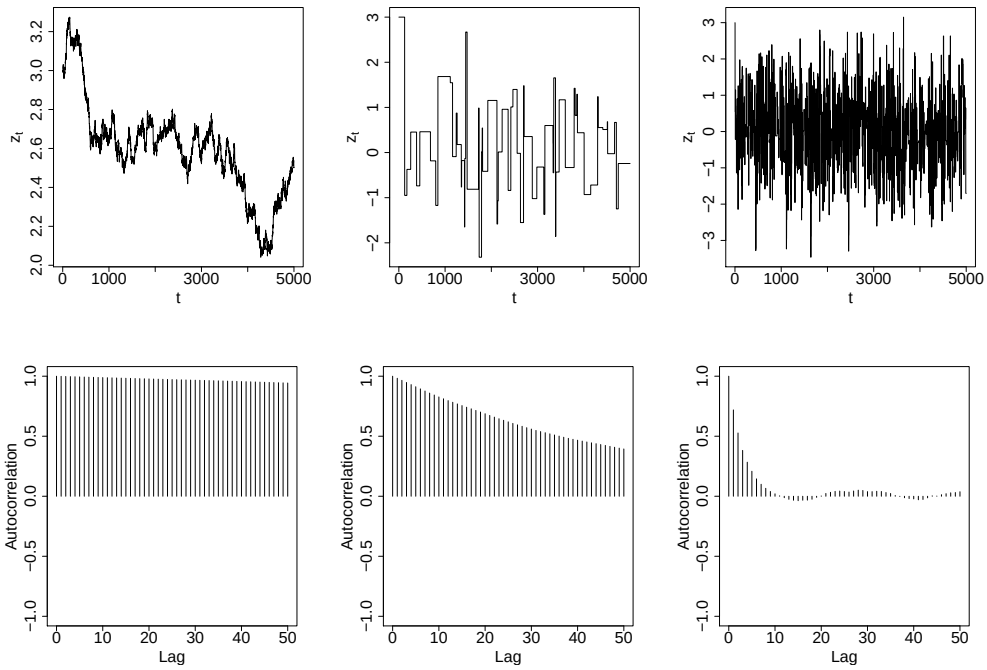

Figure 13.2: Trace plot (top plots) and ACF (bottom plots) obtained for $\sigma = 0.01$ (left plots), $\sigma = 100$ (middle plots) and $\sigma = 5$ (right plots).

The Gibbs sampler

The Gibbs sampler assumes that the distribution μ of interest is multivariate. For precisely, for an integer $d > 1$ we assume in what follows that $\mathcal{Z} = \times_{i=1}^{d} \mathcal{Z}_i$ where $\mathcal{Z}_i \subset \mathbb{R}^{k_i}$ for all i.

The introduction of the Gibbs sampler requires some additional notation. Let $z^{(i)} \in \mathcal{Z}_i$ be the ith 'block' of coordinates of $z \in \mathcal{Z}$, so that

$$z = (z^{(1)}, \ldots, z^{(d)}),$$

and let $z^{(k:l)} = (z^{(k)}, \ldots, z^{(l)})$. We denote by $z^{(-i)}$ the vector z without its ith block, that is (with obvious conventions when $i \in \{1, d\}$).

$$z^{(-i)} = (z^{(1:i-1)}, z^{(i+1:d)}).$$

Finally, for $\mu \in \mathcal{P}(\mathcal{Z})$, $i \in \{1, \ldots, d\}$ and $z \in \mathcal{Z}$, we let $\mu^{(i)}(\cdot|z^{(-i)})$ be the p.d.f. on $\mathcal{Z}_i$ defined by

$$\mu^{(i)}(\tilde{z}|z^{(-i)}) = \frac{\mu\big(z^{(1:i-1)}, \tilde{z}, z^{(i+1):d}\big)}{\int_{\mathcal{Z}_i} \mu\big(z^{(1:i-1)}, u, z^{(i+1):d}\big) \mathrm{d}u}, \quad \forall \tilde{z} \in \mathcal{Z}_i$$

In words, $\mu^{(i)}(\cdot|z^{(-i)})$ is the p.d.f. of the distribution of $Z^{(i)}$ under μ, conditional to $Z^{(-i)} = z^{(-i)}$.

Remark: The distribution $\mu^{(i)}(\cdot|z^{(-i)})$ is sometimes called the full conditional distribution of $Z^{(i)}$.

The Gibbs sampler: Algorithm

Using the above notation the Gibbs sampler on a general state space $\mathcal{Z}$ works as follows.

Gibbs sampler (Algorithm A3)

Input: $\mu \in \mathcal{P}(\mathcal{Z})$, $z_0 \in \mathcal{Z}$

Set $Z_0 = z_0$

for $t \geq 1$ **do**

  for $i = 1, \ldots, d$ **do**

    $Z_t^{(i)} \sim \mu^{(i)}(z_t^{(i)} | Z_t^{(1:i-1)}, Z_{t-1}^{(i+1:d)}) \mathrm{d}z_t^{(i)}$

  end for

end for

Remark: The fact the Markov kernel P^{Gibbs} of the Markov chain $(Z_t)_{t \geq 1}$ defined by Algorithm A3 has μ as invariant distribution is obvious.

Compared to the M-H algorithm, the main advantage of the Gibbs sampler is that it does not require to choose a proposal distribution Q; that is, implementing the Gibbs sampler only requires to specify the distribution of interest $\mu \in \mathcal{P}(\mathcal{Z})$ and a starting value $z_0 \in \mathcal{Z}$.

However:

1. The resulting Markov chain may have a large mixing time if the correlation among the different coordinates $Z \sim \mu$ is important (see the example below).
2. To implement Algorithm A3 we must be able to sample from the full conditional distribution $\mu^{(i)}(\cdot | z^{(-i)})$ for all i, which is rarely the case in practice (see below for a solution to this problem).

Gibbs sampler: A toy example

Let $\mathcal{Z} = \mathbb{R}^2$ and $\mu(z)\mathrm{d}z = \mathcal{N}_2\big(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\big)$, so that $\mu^{(i)}(z^{(i)}|z^{(-i)})\mathrm{d}z^{(i)} = \mathcal{N}_1(\rho\, z^{(-i)}, 1-\rho^2)$ for $i \in \{1,2\}$.

The trace plot and ACF for $(Z_t^{(1)})_{t\geq 1}$ obtained with the Gibbs sampler (Algorithm A3), with starting value $z_0 = (3,3)$, are shown in Figure 13.3. The results, reported for all $\rho \in \{0.98, 0.7, 0.3\}$, shows that as the correlation between the components of $Z \sim \mu(z)\mathrm{d}z$ decreases (in absolute value) the mixing time improves (i.e. reduces).

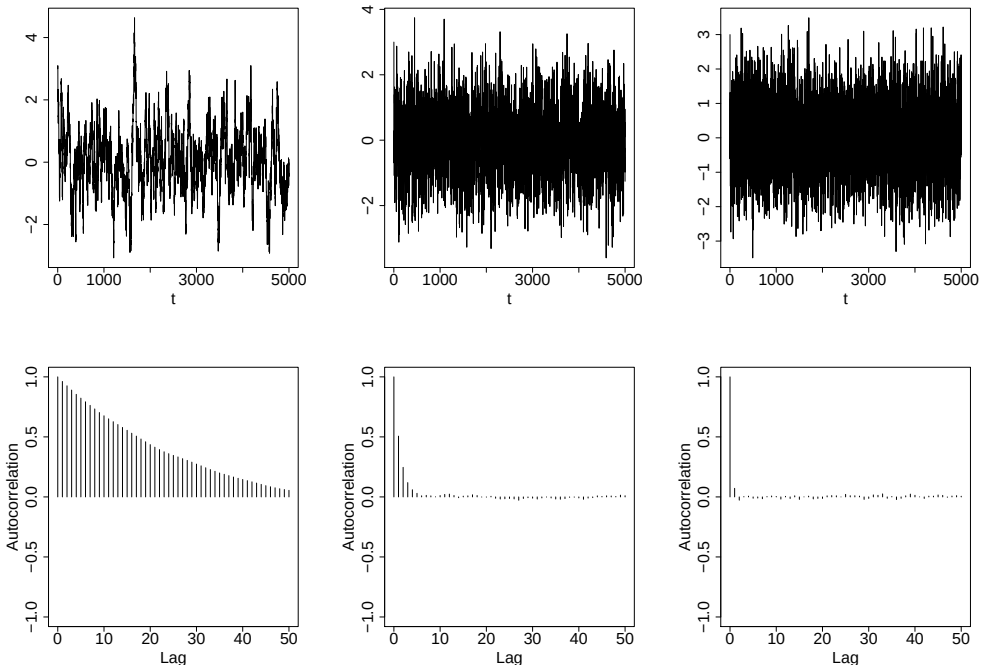

Figure 13.3: Trace plot (top plots) and ACF (bottom plots) for $(Z_t^{(1)})_{t\geq 1}$ obtained for $\rho = 0.98$ (left plots), $\rho = 0.7$ (middle plots) and $\rho = 0.3$ (right plots).

The Metropolis-within-Gibbs (MwG) algorithm

As mentioned above, the Gibbs sampler (Algorithm A3) requires to be able to sample from $\mu^{(i)}(\cdot|z^{(-i)})$ for all i, which is rarely the case in practice. Hence, we assume assume that we can simulate from $\mu^{(i)}(\cdot|z^{(-i)})$ only for $i = 1, \dots, d_1$, for some $d_1 \in \{0, \dots, d\}$.

In this context, the following modified Gibbs sampler can be used.

Hybrid Gibbs sampler (Algorithm A4)

Input: $\mu \in \mathcal{P}(\mathcal{Z})$, $z_0 \in \mathcal{Z}$

Set $Z_0 = z_0$

for $t \geq 1$ **do**

  for $i = 1, \dots, d_1$ **do**

    $Z_t^{(i)} \sim \mu^{(i)}(z_t^{(i)}|Z_t^{(1:i-1)}, Z_{t-1}^{(i+1:d)})\mathrm{d}z_t$

  end for

  for $i = d_1 + 1, \dots, d$ **do**

    $Z_t^{(i)} \sim P_{\mu^{(i)}(\cdot|Z_t^{(1:i-1)}, Z_{t-1}^{(i+1:d)})}(Z_{t-1}^{(i)}, \mathrm{d}z_t^{(i)})$

  end for

end for

Notation: We denote by P_η a transition kernel having η as invariant distribution.

In practice, $P_{\mu^{(i)}(\cdot|z^{(1:i-1)}, z_{t-1}^{(i+1:d)})}$ is often a M-H kernel, and the resulting algorithm is known as the Metropolis-within-Gibbs algorithm.

Remark: The MwG algorithm is also useful when we can sample from none of the full conditional distributions, as finding a good proposal distribution for a M-H algorithm targetting the low dimensional distribution $\mu^{(i)}(\cdot|z^{(1:i-1)}, z_{t-1}^{(i+1:d)})$ may be easier than finding a good proposal distribution for a M-H algorithm targetting μ

Invariant distribution of the Metropolis-within-Gibbs algorithm

The validity of Algorithm A4 for approximating μ relies on the following lemma.

Lemma 13.5 *Let $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2$ and $\mu(z)\mathrm{d}z \in \mathcal{P}(\mathcal{Z})$. For every $z \in \mathcal{Z}$ let $k_{z^{(2)}}(\cdot|z^{(1)})$ be a Markov kernel on $\mathcal{Z}_1$ having $\mu^{(1)}(\cdot|z^{(2)})$ as invariant distribution and let $k'_{z^{(1)}}(\cdot|z^{(2)})$ be a Markov kernel on $\mathcal{Z}_2$ having $\mu^{(2)}(\cdot|z^{(1)})\mathrm{d}\tilde{z}^{(1)}$ as invariant distribution. Then, Markov kernel K on $\mathcal{Z}$, defined by*

$$K(z,\tilde{z})\mathrm{d}\tilde{z} = k_{z^{(2)}}(\tilde{z}^{(1)}|z^{(1)})k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mathrm{d}\tilde{z}, \quad z \in \mathcal{Z}$$

has μ as invariant distribution.

Proof: For all $\tilde{z} \in \mathcal{Z}$ we have

$$\begin{aligned}
\int_{\mathcal{Z}} K(z,\tilde{z})\mu(\mathrm{d}z) &= \int_{\mathcal{Z}} k_{z^{(2)}}(\tilde{z}^{(1)}|z^{(1)})k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mu(z)\mathrm{d}z \\
&= \int_{\mathbb{R}^2} k_{z^{(2)}}(\tilde{z}^{(1)}|z^{(1)})k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mu^{(1)}(z^{(1)}|z^{(2)})\mu(z^{(2)})\mathrm{d}z \\
&= \int_{\mathcal{Z}_2} \Big(\int_{\mathcal{Z}_1} k_{z^{(2)}}(\tilde{z}^{(1)}|z^{(1)})\mu^{(1)}(z^{(1)}|z^{(2)})\mathrm{d}z^{(1)} \Big) k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mu(z^{(2)})\mathrm{d}z^{(2)} \\
&= \int_{\mathcal{Z}_2} \mu^{(1)}(\tilde{z}^{(1)}|z^{(2)})k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mu(z^{(2)})\mathrm{d}z^{(2)} \\
&= \int_{\mathcal{Z}_2} \mu(\tilde{z}^{(1)}, z^{(2)})k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mu(z^{(2)})\mathrm{d}z^{(2)} \\
&= \mu(\tilde{z}_1) \int_{\mathcal{Z}_2} \mu^{(2)}(z^{(2)}|\tilde{z}^{(1)})k'_{\tilde{z}^{(1)}}(\tilde{z}^{(2)}|z^{(2)})\mathrm{d}z^{(2)} \\
&= \mu_1(\tilde{z}_1)\mu^{(2)}(\tilde{z}^{(2)}|\tilde{z}^{(1)}) \\
&= \mu(\tilde{z}^{(1)}, \tilde{z}^{(2)}) \\
&= \mu(\tilde{z})
\end{aligned}$$

as required. □

Lemma 13.5 shows that Algorithm A4 as μ as invariant distribution when $d = 2$. This result can be extended to an arbitrary $d \geq 2$ by induction on d.

Gibbs sampler v.s. MwG algorithm: A toy example

As per above we let $\mathcal{Z} = \mathbb{R}^2$ and $\mu(z)\mathrm{d}z = \mathcal{N}_2\big(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\big)$, and our goal is to compare the performance of the Gibbs sampler and of a MwG algorithm, both starting at $z_0 \in \mathcal{Z}$, to generate a sample that approximates μ.

The Metropolis-within-Gibbs algorithm we consider is as follows.

Metropolis-within-Gibbs for the Bivariate Gaussian example (Algorithm A5)

Input: $z_0 \in \mathcal{Z}$ and $\sigma > 0$

Set $Z_0 = z_0$

for $t \geq 1$ **do**

$Z_t^{(1)} \sim \mathcal{N}_1(\rho Z_{t-1}^{(2)}, 1 - \rho^2)$

$\tilde{Z}_t^{(2)} \sim \mathcal{N}_1(Z_{t-1}^{(2)}, \sigma^2)$

Set $Z_t^{(2)} = \tilde{Z}_t^{(2)}$ with probability

$$\min\left\{1, \frac{\varphi\big(\tilde{Z}_t^{(2)}; \rho Y_t^{(1)}, 1 - \rho^2\big)}{\varphi\big(Z_{t-1}^{(2)}; \rho Z_t^{(1)}, 1 - \rho^2\big)}\right\}$$

and $Z_t^{(2)} = Z_{t-1}^{(2)}$ otherwise.

end for

Notation: $\varphi(\cdot; m, s^2)$ denotes the p.d.f. of the $\mathcal{N}_1(m, s^2)$ distribution.

For this experiment we let $z_0 = (3, 3)$, $\sigma = 2$ and $\rho = 0.9$.

Gibbs sampler v.s. MwG algorithm: An example (end)

The trace plots and ACFs for $(Z_t^{(2)})_{t\geq 1}$ obtained with the Gibbs sampler (Algorithm A3) and with the MwG algorithm (Algorithm A5) are presented in Figure 13.4.

We observe from this figure that the autocorrelations are larger for Markov chain generated by the MwG algorithm than for the one generated by the Gibbs sampler. As a result, the estimator $\hat{\mu}_T(\varphi)$ (defined in (13.9)) computed with the former algorithm will have a larger (asymptotic) variance than the one computed with the latter algorithm (see (13.10)).

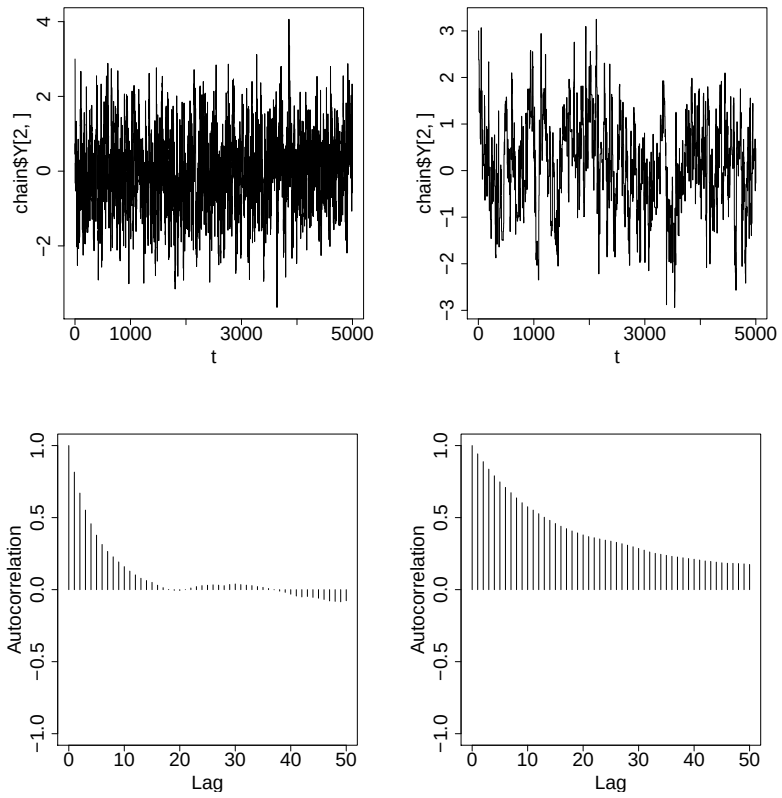

Figure 13.4: Trace plot (top plots) and ACF (bottom plots) for $(Z_t^{(2)})_{t\geq 1}$ obtained with Gibbs sampler (left plots) and with the Metropolis-within-Gibbs algorithm (right plot).

References

[1] Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

[2] Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

[3] Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[4] Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

[5] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

[6] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[7] Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339.

[8] Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

[9] Mairal, J. and Yu, B. (2012). Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*.

[10] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press Inc.

[11] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

[12] Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

[13] Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

[14] Quinonero-Candela, J. and Rasmussen, C. E. (2005). Analysis of some methods for reduced rank gaussian process regression. In *Switching and learning in feedback systems*, pages 98–127. Springer.

[15] Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

[16] Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

[17] Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367.

[18] Sande, E., Manni, C., and Speleers, H. (2020). Explicit error estimates for spline approximation of arbitrary smoothness in isogeometric analysis. *Numerische Mathematik*, 144(4):889–929.

[19] Särkkä, S. and Solin, A. (2019). *Applied stochastic differential equations*, volume 10. Cambridge University Press.

[20] Sniekers, S., van der Vaart, A., et al. (2015). Adaptive bayesian credible sets in regression with a gaussian process prior. *Electronic Journal of Statistics*, 9(2):2475–2527.

[21] Szabó, B., Van Der Vaart, A. W., van Zanten, J., et al. (2015). Frequentist coverage of adaptive nonparametric bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428.

[22] Vaart, A. v. d. and Zanten, H. v. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119.

[23] van der Vaart, A. W., van Zanten, J. H., et al. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675.

[24] van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

[25] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[26] Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.