Advanced Simulation Methods

Chapter 6 - The Metropolis-Hastings Algorithm

In this chapter we introduce the Metropolis-Hastings algorithm, which is arguably the most useful MCMC algorithm. Similarly to rejection sampling, the Metropolis-Hastings relies on a proposal distribution and the proposed values are accepted with respect to a given probability to ensure that the invariant distribution of Markov chain is the target distribution of interest. A key advantage of Metropolis-Hastings is that contrary to rejection sampling, it is possible to design "local" proposals; i.e. the proposal at iteration t can depend on the last accepted value. We will then show that it is possible to combine Metropolis-Hastings kernels to obtain more sophisticated algorithms and we will recover the Gibbs sampler as a particular case.

1 Metropolis-Hastings Algorithm

Assume that we are interested in sampling from a distribution on $\mathbb{X}$ with probability density function π. In most applications, we will have $\mathbb{X} = \mathbb{R}^d$. We introduce a proposal distribution on the space $\mathbb{X}$, with density written $q\left(x \middle| x'\right)$; i.e. for any $x' \in \mathbb{X}$ we have $q\left(x \middle| x'\right) \geq 0$ and $\int_{\mathbb{X}} q\left(x \middle| x'\right) dx = 1$. Note that the proposal distribution q can depend on x'. The Metropolis-Hastings algorithm proceeds as follows to generate a Markov chain $\left(X^{(t)}\right)_{t \geq 1}$.

Algorithm. Metropolis-Hastings. Starting from an arbitrary $X^{(1)}$, iterate for $t = 2, 3, ...$

1. Sample $X \sim q\left(\cdot \middle| X^{(t-1)}\right)$.
2. Compute
$$\alpha\left(X \middle| X^{(t-1)}\right) = \min\left(1, \frac{\pi\left(X\right) q\left(X^{(t-1)} \middle| X\right)}{\pi\left(X^{(t-1)}\right) q\left(X \middle| X^{(t-1)}\right)}\right)$$
3. With probability $\alpha\left(X \middle| X^{(t-1)}\right)$, set $X^{(t)} = X$, otherwise set $X^{(t)} = X^{(t-1)}$.

Note that the Metropolis-Hastings only requires point-wise evaluations $\pi\left(x\right)$ up to a normalizing constant, since if $\widetilde{\pi}\left(x\right) \propto \pi\left(x\right)$ then

$$\forall x \in \mathbb{X} \quad \frac{\pi\left(x\right) q\left(x^{(t-1)} \middle| x\right)}{\pi\left(x^{(t-1)}\right) q\left(x \middle| x^{(t-1)}\right)} = \frac{\widetilde{\pi}\left(x\right) q\left(x^{(t-1)} \middle| x\right)}{\widetilde{\pi}\left(x^{(t-1)}\right) q\left(x \middle| x^{(t-1)}\right)}.$$

At each iteration t, we propose a candidate from $q\left(\cdot \middle| X^{(t-1)}\right)$. The probability of accepting a candidate, given the current state is $X^{(t-1)} = x^{(t-1)}$, is given by

$$a\left(x^{(t-1)}\right) = \int_{\mathbb{X}} \alpha\left(x \middle| x^{(t-1)}\right) q\left(x \middle| x^{(t-1)}\right) dx,$$

in which case we set $X^{(t)} = X$; otherwise the chain stays where it was already, i.e. at $x^{(t-1)}$.

This algorithm generates a Markov chain $\left(X^{(t)}\right)_{t \geq 1}$. The following expression establishes the expression of the associated Markov chain kernel.

Proposition 1.1. *The transition kernel of the Metropolis-Hastings algorithm is given by*

$$K\left(x^{(t-1)}, x^{(t)}\right) = \alpha\left(x^{(t)} \middle| x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right) + \left(1 - a\left(x^{(t-1)}\right)\right) \delta_{x^{(t-1)}}\left(x^{(t)}\right)$$

where $\delta_{x^{(t-1)}}$ denotes the Dirac mass at $x^{(t-1)}$.

Proof. We have for any $A \subset \mathbb{X}$,

$$\begin{aligned}
&\mathbb{P}\left(X^{(t)} \in A \middle| X^{(t-1)} = x^{(t-1)}\right) \\
&= \mathbb{P}\left(X^{(t)} \in A, \text{proposal accepted} \middle| X^{(t-1)} = x^{(t-1)}\right) \\
&+ \mathbb{P}\left(X^{(t)} \in A, \text{proposal rejected} \middle| X^{(t-1)} = x^{(t-1)}\right) \\
&= \mathbb{P}\left(X^{(t)} \in A, \text{proposal accepted} \middle| X^{(t-1)} = x^{(t-1)}\right) \\
&+ \mathbb{P}\left(X^{(t)} \in A \middle| X^{(t-1)} = x^{(t-1)}, \text{proposal rejected}\right) \mathbb{P}\left(\text{proposal rejected} \middle| X^{(t-1)} = x^{(t-1)}\right) \\
&= \int_A \underbrace{\int_{\mathbb{X}} \delta_x\left(x^{(t)}\right) \alpha\left(x \middle| x^{(t-1)}\right) q\left(x \middle| x^{(t-1)}\right) dx}_{=\alpha\left(x^{(t)} \middle| x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right)} dx^{(t)} + \underbrace{\int_A \delta_{x^{(t-1)}}\left(x^{(t)}\right) dx^{(t)}}_{=\mathbb{I}_A\left(x^{(t)}\right)} \cdot \left(1 - a\left(x^{(t-1)}\right)\right).
\end{aligned}$$

By definition of the Markov kernel, we have also

$$\mathbb{P}\left(X^{(t)} \in A \middle| X^{(t-1)} = x^{(t-1)}\right) = \int_A K\left(x^{(t-1)}, x^{(t)}\right) dx^{(t)}$$

and the result follows. ■

2 Convergence Results

We now establish that the Metropolis-Hastings algorithm is $\pi-$reversible and admits π as invariant distribution.

Proposition 2.1. *The Metropolis-Hastings kernel is $\pi-$reversible; i.e. for any $x, y \in \mathbb{X}$ we have*

$$\pi\left(x^{(t-1)}\right) K\left(x^{(t-1)}, x^{(t)}\right) = \pi\left(x^{(t)}\right) K\left(x^{(t)}, x^{(t-1)}\right)$$

and thus it admits π as invariant distribution:

$$\int_{\mathbb{X}} \pi\left(x^{(t-1)}\right) K\left(x^{(t-1)}, x^{(t)}\right) dx^{(t-1)} = \pi\left(x^{(t)}\right).$$

Proof. We have

$$\begin{aligned}
&\pi\left(x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right) \alpha\left(x^{(t)} \middle| x^{(t-1)}\right) \\
&= \pi\left(x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right) \min\left(1, \frac{\pi\left(x^{(t)}\right) q\left(x^{(t-1)} \middle| x^{(t)}\right)}{\pi\left(x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right)}\right) \\
&= \min\left(\pi\left(x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right), \pi\left(x^{(t)}\right) q\left(x^{(t-1)} \middle| x^{(t)}\right)\right) \\
&= \pi\left(x^{(t)}\right) q\left(x^{(t-1)} \middle| x^{(t)}\right) \min\left(1, \frac{\pi\left(x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right)}{\pi\left(x^{(t)}\right) q\left(x^{(t-1)} \middle| x^{(t)}\right)}\right) \\
&= \pi\left(x^{(t)}\right) q\left(x^{(t-1)} \middle| x^{(t)}\right) \alpha\left(x^{(t-1)} \middle| x^{(t)}\right)
\end{aligned}$$

and thus

$$\begin{aligned}
\pi\left(x^{(t-1)}\right) K\left(x^{(t-1)}, x^{(t)}\right) &= \underbrace{\pi\left(x^{(t-1)}\right) q\left(x^{(t)} \middle| x^{(t-1)}\right) \alpha\left(x^{(t)} \middle| x^{(t-1)}\right)}_{=\pi\left(x^{(t)}\right) q\left(x^{(t-1)} \middle| x^{(t)}\right) \alpha\left(x^{(t-1)} \middle| x^{(t)}\right)} \\
&+ \underbrace{\pi\left(x^{(t-1)}\right)\left(1 - a\left(x^{(t-1)}\right)\right) \underbrace{\delta_{x^{(t-1)}}\left(x^{(t)}\right)}_{=0 \text{ if } x^{(t)} \neq x^{(t-1)}}}_{\left(1-a\left(x^{(t)}\right)\right) \delta_{x^{(t)}}\left(x^{(t-1)}\right)} \\
&= \pi\left(x^{(t)}\right) K\left(x^{(t)}, x^{(t-1)}\right).
\end{aligned}$$

As the kernel is π-reversible then it is π-invariant. ■

Similarly to the Gibbs sampler, establishing that the transition kernel admits π as invariant distribution is not sufficient; we also need π to be the limiting distribution. On top of reversibility, we require that the Markov chain is π-irreducible.

Example 2.1. *Reducible Metropolis-Hastings. Consider* $\mathbb{X} = \mathbb{R}$ *and* $\pi(x) = \frac{1}{2}\pi_1(x) + \frac{1}{2}\pi_2(x)$ *where* $\pi_1(x), \pi_2(x)$ *are two probability density functions such that supp* $\pi_1 = (-\infty, 0)$ *and supp* $\pi_2 = (1, \infty)$. *If we use as a proposal* $q(x'|x)$ *a uniform density* $U_{(x-\delta, x+\delta)}$ *centered around* x *then the Markov chain will not be* $\pi-$*irreducible if* $\delta < 1$. *It will be stuck either in* $(-\infty, 0)$ *or* $(1, \infty)$.

Under weak assumptions on the proposal, we can guarantee that we can reach any set of non-zero probability under π. This is for example the case if $q(x|x') > 0$ for any $x, x' \in$ supp π, in which case the chain is strongly π-irreducible. This is satisfied for example if $\mathbb{X} = \mathbb{R}^d$ and $q(x|x')$ is a normal distribution with mean x' and with a non-degenerate covariance matrix.

If the target density is bounded away from 0 and ∞ on compact sets, and if there exist $\delta, \varepsilon > 0$ such that for every $x \in \mathbb{X}$

$$|x - x'| < \delta \Rightarrow q(x|x') \geq \varepsilon,$$

then it can be shown that the chain is π-irreducible; see (Roberts & Rosenthal, 2004).

Note also that a Markov chain is aperiodic if there is a positive probability of it staying at a given state, i.e. $\mathbb{P}\left(X^{(t)} = X^{(t-1)}\right) > 0$, which is typically the case for a Metropolis-Hastings chain[1].

Theorem 2.1. *Assume the Markov chain generated by the Metropolis-Hastings sampler is* $\pi-$*irreducible then we have for any integrable function* $\phi : \mathbb{X} \to \mathbb{R}$*:*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \phi\left(X^{(i)}\right) = \int \phi(x)\,\pi(x)\,dx$$

for every starting value $X^{(1)}$.

Proof. It follows from the fact that if a Metropolis-Hastings sampler is $\pi-$irreducible then it is Harris recurrent; see Corollary 2 in (Tierney, 1994). ■

On top of this law of large number, we could also look at geometric ergodicity, central limit theorems, etc. See the references, and the references therein, for many results on the behavior of the Metropolis-Hastings algorithm.

3 Proposals for Metropolis-Hastings

So far we have described the algorithm with a generic proposal kernel q with density $q(x \mid x')$; again it means that for any x', $x \mapsto q(x \mid x')$ is a probability density function. Let us describe the main choices made in practice.

3.1 Independent Proposals

Consider the scenario where $q\left(x|x^{(t-1)}\right) = q(x)$, then the Metropolis-Hastings acceptance ratio is given by

$$\frac{\pi(x)\,q\left(x^{(t-1)}\middle|x\right)}{\pi\left(x^{(t-1)}\right)q\left(x|x^{(t-1)}\right)} = \frac{\pi(x)}{q(x)}\frac{q\left(x^{(t-1)}\right)}{\pi\left(x^{(t-1)}\right)}.$$

Contrary to accept-reject, we do not need to ensure that $\pi(x) \leq Mq(x)$ for some M. Whenever $q(x) > 0$ for any $x \in \mathbb{X}$, Theorem 2.1 holds. However, if this boundedness condition is not satisfied then convergence of the ergodic averages $\frac{1}{t}\sum_{i=1}^{t}\phi\left(X^{(i)}\right)$ towards $\int \phi(x)\,\pi(x)\,dx$ can be very slow.

[1] This condition is not met when $q\left(x^{(t)}\middle|x^{(t-1)}\right) = \pi\left(x^{(t)}\right)$, but then the chain would be made of i.i.d. samples from π, hence it would be aperiodic anyway.

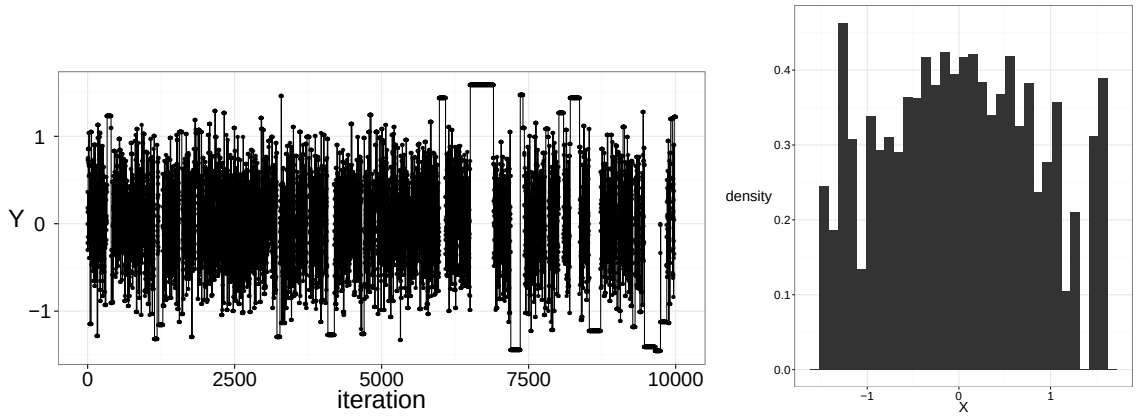

Figure 1: Independent Metropolis-Hastings output for q_1, with a standard deviation of 0.4.

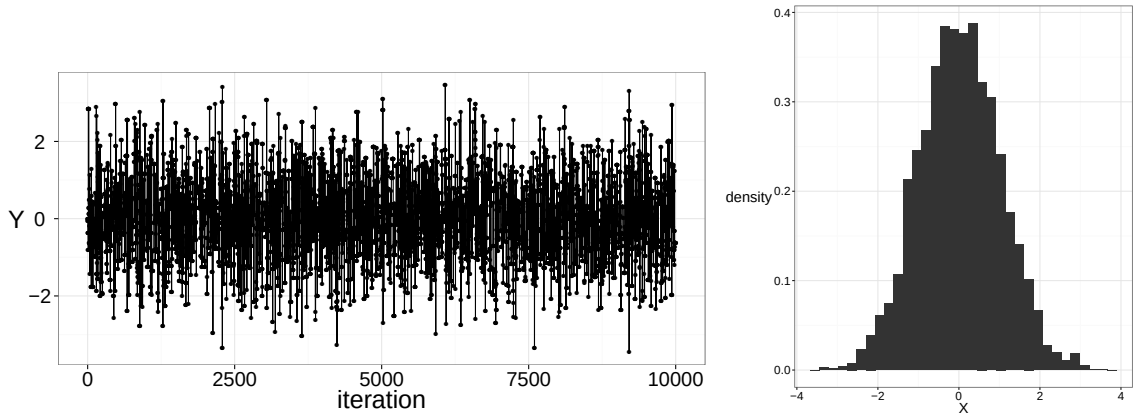

Figure 2: Independent Metropolis-Hastings output for q_2, with a standard deviation of 5.

Example 3.1. *Consider the case where* $\pi(x) = \mathcal{N}(x; 0, 1)$ *and we implement Metropolis-Hastings with two distinct independent proposals:* $q_1(x) = \mathcal{N}\left(x; 0, 0.4^2\right)$ *which is such that* $\pi(x)/q_1(x) \to \infty$ *as* $|x| \to \infty$ *and* $q_2(x) = \mathcal{N}\left(x; 0, 5^2\right)$ *which is such that* $\pi(x)/q_2(x) \leq M < \infty$*. We display in Figures 1 and 2 a trace plot of the chains of* $X^{(t)}$*, and a histogram of them, to be compared with the target distribution* $\mathcal{N}(x; 0, 1)$*. When using* q_1*, the chain gets stuck in the tails of the target as the value of* $\pi\left(x^{(t-1)}\right)/q_1\left(x^{(t-1)}\right)$ *at the current state is then large; thus the acceptance probability of a new proposal is low.*

3.2 Random Walk Proposals

Consider now the scenario where we use a random walk proposal, that is

$$X = X^{(t-1)} + W$$

where $W \sim g$, g being a symmetric proposal $g(-w) = g(w)$; e.g. g could be a zero-mean normal of covariance Σ or t-student. In this case, we have

$$q\left(x \middle| x^{(t-1)}\right) = g\left(x - x^{(t-1)}\right) = g\left(-x + x^{(t-1)}\right) = q\left(x^{(t-1)} \middle| x\right),$$

so the Metropolis-Hastings acceptance ratio becomes

$$\frac{\pi(x)\, q\left(x^{(t-1)} \middle| x\right)}{\pi\left(x^{(t-1)}\right) q\left(x \middle| x^{(t-1)}\right)} = \frac{\pi(x)}{\pi\left(x^{(t-1)}\right)}.$$

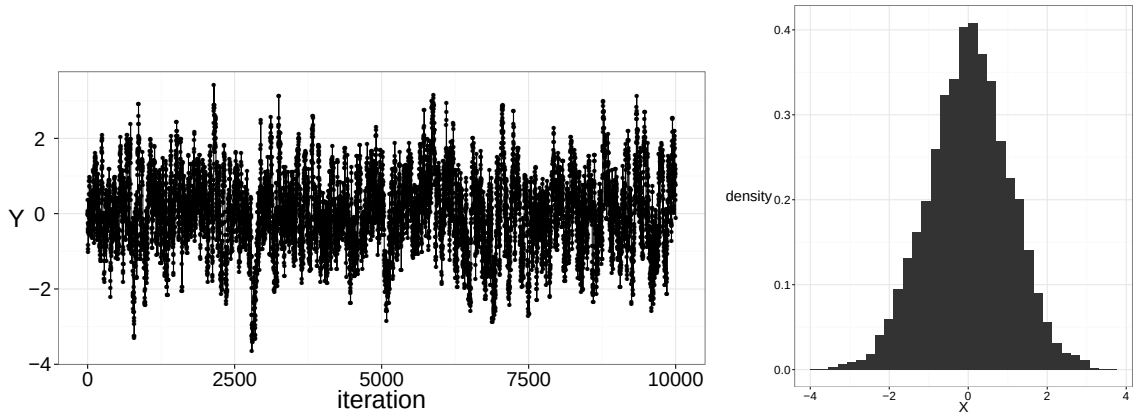

Figure 3: Random Walk Metropolis-Hastings output for q_1, with a standard deviation of 0.4.

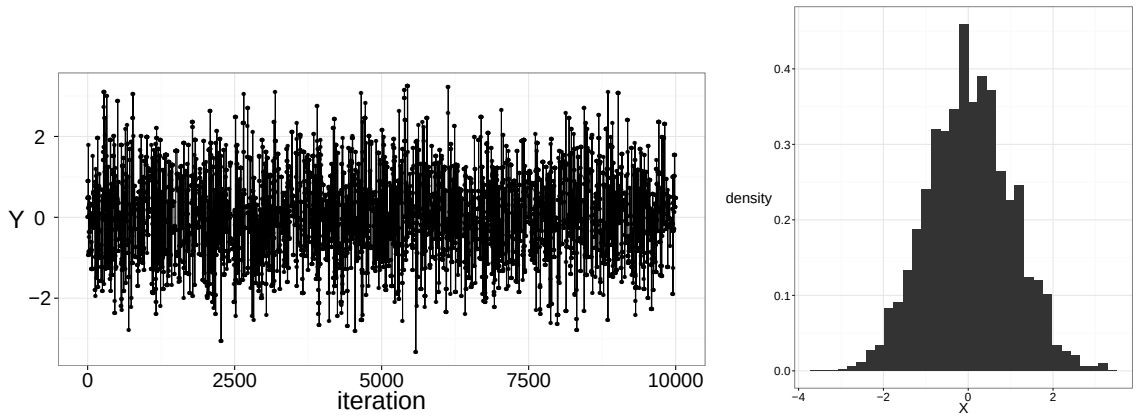

Figure 4: Random Walk Metropolis-Hastings output for q_2, with a standard deviation of 5.

Example 3.2. *Consider the case where* $\pi(x) = \mathcal{N}(x; 0, 1)$ *and we implement Metropolis-Hastings with two random walk proposals:* $q_1\left(x \middle| x^{(t-1)}\right) = \mathcal{N}\left(x; x^{(t-1)}, 0.4^2\right)$, $q_2\left(x \middle| x^{(t-1)}\right) = \mathcal{N}\left(x; x^{(t-1)}, 5^2\right)$. *We display in Figures 3-4 a trace plot of the chains of* $X^{(t)}$*, and a histogram of them, to be compared with the target distribution* $\mathcal{N}(x; 0, 1)$.

Heavy tails distribution g can generally prevent the chain from getting "trapped" in local modes of the target distribution. If the variance of the random walk proposal is small compared to the "width" of the target distribution, then it takes numerous iterations to explore the whole support of the target. It is tempting to adapt the variance of the increments given the simulation output... unfortunately this breaks the Markov property and biases results if one is not careful; this technique is usually called "adaptive MCMC" and has been the topic of many recent articles in the MCMC literature.

3.3 Local Optimization Proposals

It is also possible to use "sophisticated" proposals. For example, the so-called Langevin algorithm relies on

$$X = X^{(t-1)} + \frac{\sigma}{2} \nabla \log \pi|_{X^{(t-1)}} + \sigma W$$

where $W \sim \mathcal{N}(0, I_d)$, and $\nabla \log \pi|_x$ refers to the gradient of the log target density function evaluated at x. The rationale of this proposal distribution is to "follow the slope" of the target distribution, in order to propose points that are more likely to have higher target density values that the current point of the

chain. Then the Metropolis-Hastings acceptance ratio is

$$\frac{\pi(x)\, q\left(x^{(t-1)}\middle|\, x\right)}{\pi\left(x^{(t-1)}\right) q\left(x\middle|\, x^{(t-1)}\right)} = \frac{\pi(x)}{\pi\left(x^{(t-1)}\right)} \frac{\exp\left\{-\left\|x^{(t-1)} - x - \frac{\sigma}{2}\, \nabla \log \pi|_x\right\|^2 / 2\sigma^2\right\}}{\exp\left\{-\left\|x - x^{(t-1)} - \frac{\sigma}{2}\, \nabla \log \pi|_{x^{(t-1)}}\right\|^2 / 2\sigma^2\right\}}.$$

Generally speaking, we can use $q\left(x\middle|\, x^{(t-1)}\right) = g\left(x; \varphi\left(x^{(t-1)}\right)\right)$ where g is any distribution on $\mathbb{X}$ of parameters $\varphi\left(x^{(t-1)}\right)$ where φ is a deterministic mapping; say $\varphi\left(x^{(t-1)}\right)$ could correspond to a mean vector and covariance matrix for g normal. It is important to realize that we do not need to have an explicit form for φ; e.g. φ could be given by the output of a numerical algorithm such as a local optimization procedure. We will have the correct invariant distribution as long as we use the following Metropolis-Hastings acceptance ratio

$$\frac{\pi(x)\, q\left(x^{(t-1)}\middle|\, x\right)}{\pi\left(x^{(t-1)}\right) q\left(x\middle|\, x^{(t-1)}\right)} = \frac{\pi(x)\, g\left(x^{(t-1)}; \varphi(x)\right)}{\pi\left(x^{(t-1)}\right) g\left(x; \varphi\left(x^{(t-1)}\right)\right)}.$$

4 Applications

4.1 Application to Bayesian Logistic Regression

In 1986, the Challenger shuttle exploded, the explosion being the result of an O-ring failure. It was believed to be a result of a cold weather at the departure time: 31°F. We can have access to the data of 23 previous flights which give for flight i: temperature at flight time x_i, and $y_i = 1$ failure and zero otherwise (Robert & Casella, p. 281-283).

We can model the binary data using a simple logistic regression model

$$\mathbb{P}(Y = 1|\, x) = 1 - \mathbb{P}(Y = 0|\, x) = \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)}.$$

Instead of finding the maximum likelihood estimate (MLE) of (α, β) using iterative re-weighted least squares, we can follow here a Bayesian approach and assign the following prior

$$\begin{aligned} p(\alpha, \beta) &= p(\alpha|\, \beta)\, p(\beta) \\ &\propto b \exp(\alpha) \exp(-b \exp(\alpha)) \end{aligned}$$

that is exponential prior on $\exp(\alpha)$ of hyper-parameter b and a flat prior on ; b is selected as the data-dependent prior such that $\mathbb{E}(\alpha) = \widehat{\alpha}_{MLE}$. We are interested in sampling from the posterior

$$p(\alpha, \beta|\, y_1, ..., y_n) \propto p(\alpha, \beta) \prod_{i=1}^{n} \frac{\{\exp(\alpha + x_i\beta)\}^{y_i}}{1 + \exp(\alpha + x_i\beta)}$$

which can be evaluated point-wise, up to a normalizing constant. As a simple proposal for Metropolis-Hastings, we could use the independent proposal

$$q\left((\alpha, \beta)|\, \left(\alpha^{(t-1)}, \beta^{(t-1)}\right)\right) = p(\alpha|\, b)\, \mathcal{N}\left(\beta; \widehat{\beta}_{MLE}, \widehat{\sigma}^2_{\beta}\right)$$

where $\widehat{\sigma}^2_{\beta}$ is the variance associated to the MLE.

4.2 Application to Bayesian Probit Regression

We can also revisit the Bayesian probit regression model discussed in the previous chapter; see also (Marin & Robert, 2007). Remember that in this context Bayesian inference relies on

$$p(\beta|\, y_1, ..., y_n) \propto p(\beta) \prod_{i=1}^{n} \Phi\left(x_i^T \beta\right)^{y_i} \left(1 - \Phi\left(x_i^T \beta\right)\right)^{1 - y_i}.$$

In the previous chapter, we introduced latent variables in order to apply the Gibbs sampler. Here we can bypass the introduction of latent variables, and use a Metropolis-Hastings to sample from $p(\beta|\, y_1, ..., y_n)$ using a random walk proposal with normal increment of covariance matrix $\tau^2 \widehat{\Sigma}$ where $\widehat{\Sigma}^{-1}_{i,j} = \left[-\frac{\partial^2 \log p(\beta|y_1, ..., y_n)}{\partial \beta_i \partial \beta_j}\right]\Big|_{\widehat{\beta}_{\mathrm{MLE}}}$.

5 Combining MCMC kernels

In numerous practical scenarios, we have to deal with high-dimensional target distributions π, and it is difficult to find a "good" proposal for the Metropolis-Hastings algorithm; that is, one that leads to enough candidates being accepted, and simultaneously, efficient exploration of the state space; note that a random walk proposal distribution with a very small standard deviation would always lead to high acceptance ratios, as long as the target density is continuous; however it might not allow an efficient exploration of the state space. It can be helpful to use not one single MCMC kernel, but several of them, and to combine them. There are two ways of combining transition kernels, either through cycles or mixtures.

Lemma 5.1. *Assume $\{K_i\}_{i=1,...,p}$ are Markov transition kernels such that K_i is π-invariant for any $i = 1, ..., p$ then the cycle kernel*

$$K_1 K_2 \cdots K_p (x, x') := \int K_1 (x, x_2) K_2 (x_2, x_3) \cdots K_p (x_p, x') \, dx_2 \cdots dx_p$$

is π-invariant.

Proof. We have

$$\begin{aligned}
&\int \pi (x_1) K_1 K_2 \cdots K_p (x_1, x_{p+1}) \, dx_1 \cdots dx_p \\
&= \int \cdots \int \underbrace{\underbrace{\int \pi (x_1) K_1 (x_1, x_2) \, dx_1}_{\pi(x_2)} K_2 (x_2, x_3) \, dx_2}_{\pi(x_3)} \cdots K_p (x_p, x_{p+1}) \, dx_p \\
&= \pi (x_{p+1}) .
\end{aligned}$$

Hence the result follows. ■

Lemma 5.2. *Assume $\{K_i\}_{i=1,...,p}$ are Markov transition kernels such that K_i is π-invariant for any $i = 1, ..., p$ then the mixture kernel*

$$\sum_{i=1}^{p} \alpha_i K_i (x, x')$$

is π-invariant for any $\alpha_i \geq 0, \sum_{i=1}^{p} \alpha_i = 1$.

The proof is left as an exercise. This result can be used to propose algorithms which, similarly to Gibbs sampling, partition the state x into $x_1, x_2, ..., x_d$ and update component x_i conditional upon $x_{-i} := (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$. Each of the component x_i is updated using a Metropolis-Hastings algorithm of proposal q_i. If we cycle through all the components deterministically then the resulting algorithm proceeds as follows.

Algorithm. Systematic scan Metropolis-Hastings sampler. Let $\left(X_1^{(1)}, ..., X_d^{(1)}\right)$ be the initial state then iterate for $t = 2, 3, ...$

- For component 1,
  1. sample $X_1 \sim q_1 \left(\cdot | \, X_1^{(t-1)}, X_2^{(t-1)}, ..., X_d^{(t-1)} \right)$.
  2. Compute

$$\alpha_1 = \min \left(1, \frac{\pi \left(X_1, X_2^{(t-1)}, ..., X_d^{(t-1)} \right) q \left(X_1^{(t-1)} \middle| \, X_1, X_2^{(t-1)}, ..., X_d^{(t-1)} \right)}{\pi \left(X_1^{(t-1)}, X_2^{(t-1)}, ..., X_d^{(t-1)} \right) q \left(X_1 | \, X_1^{(t-1)}, X_2^{(t-1)}, ..., X_d^{(t-1)} \right)} \right)$$

  3. With probability α_1, set $X_1^{(t)} = X_1$, otherwise set $X_1^{(t)} = X_1^{(t-1)}$.

...

- For component j,

  1. Sample $X_j \sim q_j\left(\cdot|\, X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j^{(t-1)}, ..., X_d^{(t-1)}\right)$.
  2. Compute

$$\alpha_j = \min\left(1, \frac{\pi\left(X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)}\right) q\left(X_j^{(t-1)} \middle|\, X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)}\right)}{\pi\left(X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j^{(t-1)}, ..., X_d^{(t-1)}\right) q\left(X_j|\, X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j^{(t-1)}, ..., X_d^{(t-1)}\right)}\right)$$

  3. With probability α_j, set $X_j^{(t)} = X_j$, otherwise set $X_j^{(t)} = X_j^{(t-1)}$.

...

- For component d

  1. Sample $X_d \sim q_d\left(\cdot|\, X_1^{(t)}, ..., X_{d-1}^{(t)}, X_d^{(t-1)}\right)$.
  2. Compute

$$\alpha_d = \min\left(1, \frac{\pi\left(X_1^{(t)}, ..., X_{d-1}^{(t)}, X_d\right) q\left(X_d^{(t-1)} \middle|\, X_1^{(t)}, ..., X_{d-1}^{(t)}, X_d\right)}{\pi\left(X_1^{(t)}, ..., X_{d-1}^{(t)}, X_d^{(t-1)}\right) q\left(X_d|\, X_1^{(t)}, ..., X_{d-1}^{(t)}, X_d^{(t-1)}\right)}\right)$$

  3. With probability α_d, set $X_d^{(t)} = X_d$, otherwise set $X_d^{(t)} = X_d^{(t-1)}$.

The random scan algorithm proceeds as follows.

Algorithm. Random scan Metropolis-Hastings sampler. Let $\left(X_1^{(1)}, ..., X_d^{(1)}\right)$ be the initial state then iterate for $t = 2, 3, ...$

1. Sample an index J from a distribution on $\{1, ..., d\}$ (typically uniform).

   (a) Sample $X_J \sim q_J\left(\cdot|\, X_1^{(t-1)}, ..., X_d^{(t-1)}\right)$.

   (b) Compute

$$\alpha_J = \min\left(1, \frac{\pi\left(X_1^{(t-1)}, ..., X_J, ..., X_d^{(t-1)}\right) q\left(X_J^{(t-1)} \middle|\, X_1^{(t-1)}, ..., , X_J, ..., X_d^{(t-1)}\right)}{\pi\left(X_1^{(t-1)}, ..., X_d^{(t-1)}\right) q\left(X_J|\, X_1^{(t-1)}, ..., X_d^{(t-1)}\right)}\right)$$

   (c) With probability α_J, set $X_J^{(t)} = X_J$, otherwise set $X_J^{(t)} = X_J^{(t-1)}$.

2. Set $X_{-J}^{(t)} := X_{-J}^{(t-1)}$.

One can show that both algorithms admit π as invariant distribution. Note that in this scenario it is more difficult to establish π-irreducibility as none of the individual kernels are π-irreducible. The next proposition establishes the connection between Gibbs sampling and these compositions of Metropolis-Hastings kernels.

Proposition 5.1. *Systematic Gibbs sampling corresponds to a composition of Metropolis-Hastings kernels where for any $j = 1, ..., d$*

$$q\left(x_j' \middle|\, x_1, ..., x_d\right) := \pi_{X_j|X_{-j}}\left(x_j' \middle|\, x_{-j}\right).$$

Proof. It is sufficient to establish that the acceptance rate α_j is equal to 1, in order to prove this result. We have

$$\frac{\pi\left(x_1,...,x_{j-1},x'_j,x_{j+1},...,x_d\right)q\left(x_j\middle|\, x_1,...,x_{j-1},x'_j,x_{j+1},...,x_d\right)}{\pi\left(x_1,...,x_d\right)q\left(x'_j\middle|\, x_1,...,x_d\right)}$$
$$=\frac{\pi\left(x_1,...,x_{j-1},x'_j,x_{j+1},...,x_d\right)\pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)}{\pi\left(x_1,...,x_d\right)\pi_{X_j|X_{-j}}\left(x'_j\middle|\, x_{-j}\right)}$$

but

$$\pi\left(x_1,...,x_{j-1},x'_j,x_{j+1},...,x_d\right)=\pi\left(x_{-j}\right)\pi_{X_j|X_{-j}}\left(x'_j\middle|\, x_{-j}\right),$$
$$\pi\left(x_1,...,x_d\right)=\pi\left(x_{-j}\right)\pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)$$

so

$$\frac{\pi\left(x_1,...,x_{j-1},x'_j,x_{j+1},...,x_d\right)\pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)}{\pi\left(x_1,...,x_d\right)\pi_{X_j|X_{-j}}\left(x'_j\middle|\, x_{-j}\right)}$$
$$=\frac{\pi\left(x_{-j}\right)\pi_{X_j|X_{-j}}\left(x'_j\middle|\, x_{-j}\right)\pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)}{\pi\left(x_{-j}\right)\pi_{X_j|X_{-j}}\left(x_j\middle|\, x_{-j}\right)\pi_{X_j|X_{-j}}\left(x'_j\middle|\, x_{-j}\right)}=1.$$

So the result follows. ■

References

[1] C.P. Robert & G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[2] G.O. Roberts & J.S. Rosenthal, General State-Space Markov chains and MCMC Algorithms, *Probability Surveys*, vol. 4, pp. 20-71, 2004.

[3] L. Tierney, Markov chains for exploring posterior distributions, *Annals of Statistics*, vol. 22, pp. 1701-1762, 1994.