BACHELORARBEIT

Titel der Bachelorarbeit

Markov Chain theory and the Metropolis-Hastings algorithm

Verfasser

Lukas Peham

angestrebter akademischer Grad

Bachelor of Science (BSc.)

Wien, im Monat Februar 2020

Studienkennzahl lt. Studienblatt:	A 033621
Studienrichtung lt. Studienblatt:	Mathematik
Betreuer:	Assoz. Prof. Dr. Martin Ehler

Abstract

This thesis presents important results of Markov chain theory and connects them to the Metropolis Hastings algorithm, which is widely used to approximate an unknown distribution. We develop Markov chain theory for finite state spaces and show that under mild constraints (irreducible and aperiodic) a Markov chain has a unique stationary distribution towards every initial distribution converges at an exponential rate. After a brief explanation how these results transfer to Euclidean state spaces, we define the Metropolis Hastings algorithm and give a proof that the generated chain converges towards the chosen target distribution. The thesis concludes with approximations of two mixed normal distributions and the approximation of the posterior of a Bayesian linear regression model.

Contents

1 Introduction

The purpose of this thesis is to present the Metropolis Hastings algorithm, which is a well known Markov Chain Monte Carlo sampling method. This algorithm enables one to draw samples from a distribution from which it is not possible to directly draw samples. In Bayesian statistics these situations do naturally occur and therefore the Metropolis Hastings algorithm is a widely used tool in this field.

The Metropolis Hastings algorithm strongly builds on results from Markov chain theory. In order to enable us to properly introduce the algorithm we first need to gain a solid understanding of Markov chains. Therefore we carefully develop the theory of homogeneous Markov chains on finite state spaces. A great resource on Markov theory on general state spaces is [5] by Martin Hairer. We restrict ourselves to finite state spaces for the following two reasons. First, it is more intuitive and less technical to lay out the theory in this special case. Second, all the results do transfer to the general case, on which the Metropolis Hastings algorithm relies on. We will derive that an irreducible and aperiodic Markov chain converges towards its stationary distribution at an exponential rate. This convergence theorem is a major building block of the Metropolis Hastings algorithm, because the central idea of the algorithm is to turn this convergence result upside down. For a given but unknown distribution, called target distribution, the algorithm constructs a Markov chain which converges towards this unknown distribution.

The outline of the thesis looks as follows. In Section 2 we briefly introduce Monte Carlos simulation. In Section 3 we lay out the theory of finite, homogeneous Markov chains. We show that an irreducible Markov chain has a unique stationary distribution and derive the above mentioned convergence theorem. In Section 4 we define the Metropolis Hastings algorithm and proof that its generated chain does converge towards the target distribution. Then we illustrate the algorithm by sampling from two mixed normal distributions and finish with applying the algorithm to sample the posterior distribution of a simple Bayesian linear regression model.

2 Monte Carlo Simulation

The thesis starts with a brief introduction to Monte Carlo methods. In Section 4 we will introduce the Metropolis Hastings algorithm, which is a Markov Chain Monte Carlo sampling algorithm. Hence a basic understanding of Monte Carlo sampling is therefore required.

A great resource on Monte Carlo algorithms is the book [11] on which this chapter is based on.

2.1 Monte Carlo methods

Monte Carlo Simulation methods are a big class of algorithms, which at their core use a random number generator (RNG). A random number generator enables one to draw samples from some set (e.g. $[0, 1]$ or $\{0, 1, ..., N\}$). A sample of an ideal random number generator can be associated to the properties of a realisation of an independent sequence of uniformly distributed random variables. Loosely speaking, a random number generator tries to mimic an independent and uniformly distributed sequence of random variables. This is still an issue in computer science and one needs to be aware, that the outcome of a Monte Carlo algorithm depends on the quality of the used random number generator.

Formally a Monte Carlo algorithm is a sequence of computational operations, which gets a finite sequence of real numbers and functions as input and produces a finite sequence of real numbers as output. The feasible operations are:

- the four basic arithmetic operations
- logical comparison of real numbers
- evaluation of functions
- random number generator

Notice that deterministic algorithms are defined in the same way, except that they do not include a random number generator. Hence the set of deterministic algorithms is a real subset of Monte Carlo algorithms. It is interesting that there are several problems for which Monte Carlo algorithms are superior to any deterministic algorithm, given that the used random number generator is sufficiently good. E.g. numerical integration of high dimensional functions, [11, Chapter 7].

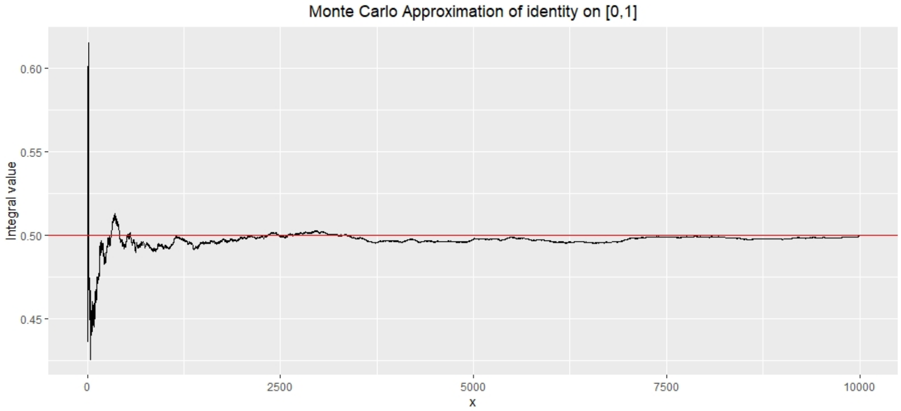
2.2 Example

Now let´s look at a simple, but instructive example of a Monte Carlo algorithm. The approximation of the integral of the identity on $[0, 1]$.

$$\int_0^1 xdx \approx \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$(x_j)_{j=1}^n \, random \, sample \, from \, [0,1]$$

The following figure shows how the Monte Carlo approximation of the integral evolves for increasing n up to $n = 10k$.



From the graph we see that the approximation is getting quite close to the true value $\frac{1}{2}$ as n gets close to $10k$. Obviously one would not use such an algorithm to approximate a simple integral. But the approximation properties of this algorithm still hold for complicated functions in high dimensional Euclidean spaces, which are impossible to handle analytically. In many of such cases, Monte Carlo methods do actually better than any deterministic algorithm, [11, Chapter 7].

3 Markov Chain Theory

This chapter presents central results of Markov chain theory, which will be used in the final Section 4 on the Metropolis Hastings algorithm. Section 3.1 lays out the basics about Markov chain theory. We start with a short recap about probability spaces, then give the definition of a Markov chain, show an example and finish with the concepts of irreducibility and stationary distributions. Sections 3.2 and 3.3 provide proofs for the unique existence of a stationary distribution. In Section 3.4 we present a very important convergence theorem and we finish with two short sections about general state spaces 3.5 and reversibility 3.6. This chapter is based on [8] and [3].

3.1 Introduction

Before starting with Markov chain theory, it is necessary to lay out basic concepts from probability theory on which the study of Markov chains will be based on. For this purpose it is not necessary to rely on measure theory, only focusing on probability theory on finite sets is enough.

Definition 3.1. A *finite probability space* is a triple $\{\Omega, \mathcal{P}(\Omega), p\}$.

With Ω being a non-empty finite set, $|\Omega| = n$ and $p \in \mathbb{R}^n$ a row vector with $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. The vector p is called the *distribution* or *probability mass function* of the probability space. The power set $\mathcal{P}(\Omega)$ accounts for all the possible events that can occur.

Definition 3.2. Let $\{\Omega, \mathcal{P}(\Omega), p\}$ be a finite probability space. Then any function $X : \Omega \to \mathbb{R}$ is called a *random variable* and the following sum

$$\mathrm{E}[X] = \sum_{x \in \Omega} p(x) X(x)$$

is called the *expected value* of the random variable X.

In finite probability theory one does not need to rely on measure theory to introduce random variables, because any function from Ω to $\mathbb{R}$ is measurable since any pre image of any Borel set is in $\mathcal{P}(\Omega)$.

Definition 3.3. A *stochastic process* in discrete time is a sequence of random variables $(X_t)_{t \in \mathbb{N}}$ onto some probability space Ω, which is referred to as the state space of the process.

Definition 3.4. A $n \times n$ Matrix Q with non negative entries which fulfills $\sum_{j=1}^n [Q]_{ij} = 1 \; \forall \, i$ is called a *stochastic Matrix.*

Proposition 3.1. *(1) Q is a stochastic matrix if and only if each row of Q is a distribution. (2) For stochastic matrices Q and R, their product QR is again a stochastic matrix.*

Proof. By a straight forward computation. □

These are the fundamentals on which the further study of Markov chains will be based on. Given all these definitions it is now possible to introduce Markov chains and start to investigate their properties.

Definition 3.5. A *finite homogeneous Markov Chain* is a stochastic process with a finite state space Ω which fulfills the following property:

$$\mathbf{P}(X_t = x_t \,|\, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \mathbf{P}(X_t = x_t \,|\, X_{t-1} = x_{t-1}) \qquad (3.1)$$

If the conditional distribution $\mathbf{P}(.|.)$ is time invariant, the Markov chain is called *homogeneous.*

Property (3.1) is called the *Markov property* and $\mathbf{P}(.|.)$ gives the probability for switching from a state x in X_t to some state y in X_{t+1}. A Markov chain is a

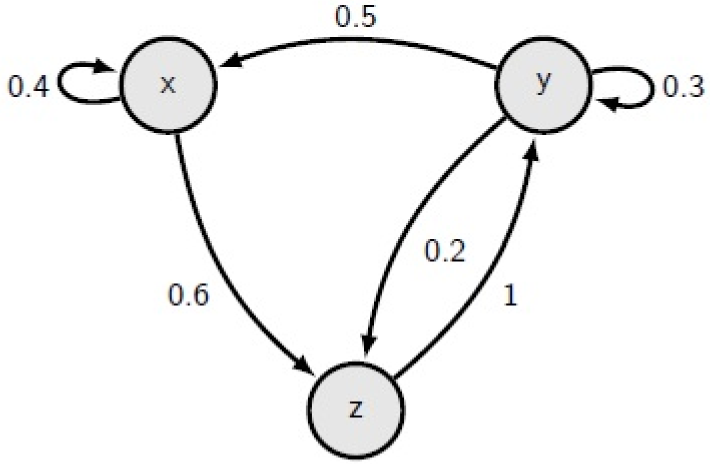


Figure 1: An example Markov chain with three states x,y and z

sequence of randomly generated states where the next state of the chain only depends on the current state, but the history of the chain is irrelevant.

The reason for the restriction to a finite state space Ω is that it makes a first start into the field of Markov chain theory less technical and more intuitive. Since finite probability spaces are easier to handle than probability spaces in the sense of measure theory. However, all the results we derive do transfer to the case of general state spaces. This will be briefly discussed in Section 3.5.

Remark. From now onward we use the term Markov chain as an equivalent for finite, homogeneous Markov chain.

Notice that for each Markov chain there exists a corresponding stochastic Matrix Q, which fully describes it. Due to the fact that Ω is finite, one can think of Ω being ordered. Take $x, y \in \Omega$ where x is the i-th and y is the j-th element in Ω. Set $[Q]_{ij} := \mathbf{P}(X_{t+1} = y|X_t = x)$ and notice that $[Q]_{ij} \geq 0 \; \forall i, j$ and that the rows of Q sum up to 1 because $\mathbf{P}$ defines a conditional probability distribution. Further notice that $\mathbf{P}(X_t = y|X_0 = x) = [Q^t]_{ij}$. The exact order of Ω does not matter as long as it is fixed and therefore we notate the $\mathbf{P}(X_t = x, X_{t-1} = y)$ element in Q with $Q(x, y)$. The reason to use this notation is that it can be used in the continuous case as well.

Remark. In order to save space, if $X_0 = x$ is given we will notate the conditional distribution on $X_0 = x$ with $\mathbf{P}_x := \mathbf{P}(.|X_0 = x)$.

There are several ways to illustrate a Markov chain. An instructive one is to use a state diagram like in Figure 1 from [3, page 2]. In such a diagram, each state is noted as a circle and the arrows indicate the probability of getting from a state to another state in one time step. E.g. in the diagram of Figure 1, the probability to move from state x to state z is 0.6. The same Markov chain can also comfortably be displayed with it´s transition matrix Q, where the sates are in the order x, y, z.

$$Q = \begin{pmatrix} 0.4 & 0 & 0.6 \\ 0.5 & 0.3 & 0.2 \\ 0 & 1 & 0 \end{pmatrix}$$

One of the first things that come to mind is, "how do distributions evolve over time in a Markov process? Will a Markov process settle down to some distribution and if so is this distribution unique? And first of all in which sense can the idea of settling down be formalised? Some more definitions are needed to address these questions.

Definition 3.6. Given a Markov chain X_t with a corresponding transformation matrix Q. A distribution π on Ω for which $\pi Q = \pi$ holds is called a *stationary distribution.*

Definition 3.7. Given a Markov chain X_t with transition matrix Q. A distribution π of Ω is called *limiting distribution* if for all initial distributions p^0

$$\lim_{t\to\infty} p^0 Q^t = \pi \quad \forall p^0$$

holds.

Proposition 3.2. *Let X_t be a Markov chain and π a limiting distribution. Then π is a stationary distribution*

Proof. A straight forward computation shows, that

$$\pi = \lim_{t\to\infty} p^0 Q^t = \lim_{t\to\infty} p^0 Q^{t+1} = \Big(\lim_{t\to\infty} p^0 Q^t\Big) Q = \pi Q$$

holds. □

Definition 3.8. A Markov chain is called *irreducible*, if

$$\forall\, z, w \in \Omega: \ \exists s \geq 1: \ such\, that\, Q^s(z, w) > 0.$$

Intuitively speaking in an irreducible Markov process it is possible to get from any state to any other state within a finite amount of time steps. On the other hand it is not possible that the process will stay in a certain state for ever. Irreducibility does not seem to be an enormous constraint, but somehow surprisingly it already guarantees the unique existence of a stationary distribution. These major result is shown in the next two subsections.

3.2 Existence of stationary distribution

First we show the existence of a stationary distribution for irreducible Markov chains by explicitly constructing one. Therefor we need the following result about hitting times. On average, we expect to get from any state to any other state in finite time.

Definition 3.9. Let X_t be a Markov chain. The *hitting time* for a state $x \in \Omega$ is the first instance of hitting the state x, notated by

$$\tau_x = \min_{t \geq 0}(X_t = x).$$

When we want the hitting time to be strictly positive, we notate it by

$$\tau_x^+ = \min_{t > 0}(X_t = x).$$

Remark. If $X_0 = x$ then τ_x^+ is called the *first return time* to x. Also notice that τ_x and τ_x^+ are random variables. Hence we can talk about their Expected value E and E_x if $X_0 = x$ is given.

Lemma 3.3. *For any $x, y \in \Omega$ of an irreducible Markov chain X_t with transition matrix Q, it holds that $\mathrm{E}_x[\tau_y^+] < \infty$.*

Proof. Since the Markov chain is irreducible, we know that for any $z, w \in \Omega$ there exists some $s \geq 1$ such that $Q^s(z, w) > 0$. Let S be the set of all these numbers. Now we can define the maximum amount of time steps to switch between any states as

$$r := \max_{s \in S} \left(Q^s(z, w) > 0,\ z, w \in \Omega\right).$$

Since Ω is finite, we simply take the maximum from a finite set of integers, hence r is finite. Now we take the smallest transition probability

$$\epsilon := \min_{0 < s \leq r, z, w \in \Omega} Q^s(z, w)$$

within a period of length r. By definition $\epsilon > 0$. For all $z, w \in \Omega$ there exists an s with $0 < s \leq r$ such that $Q^s(z, w) \geq \epsilon$. Hence given X_t, the probability of the chain going to a state y between times t and $t + r$ is at least ϵ. Therefore

$$\mathbf{P}(X_s \neq y) \leq \epsilon \quad \forall s : t < s \leq t + r \tag{3.2}$$

holds.

Now choose $k \in \mathbb{N}, k > 0$ and notice that

$$\mathbf{P}_x(\tau_y^+ > kr) = \mathbf{P}_x(X_t \neq y \; \forall t: \; 0 < t \leq (k-1)r)\mathbf{P}_x(X_t \neq y \; \forall t: \; (k-1)r < t \leq kr) \tag{3.3}$$

holds.

Applying the above inequality (3.2) iteratively on (3.3) gives

$$\mathbf{P}_x(\tau_y^+ > kr) \leq (1-\epsilon)^k. \tag{3.4}$$

By definition of the expected value and reordering of the partial sums we get

$$\mathrm{E}_x[\tau_y^+] = \sum_{t=0}^{\infty} t\mathbf{P}_x(\tau_y^+ = t) = \sum_{t=0}^{\infty} \mathbf{P}_x(\tau_y^+ > t). \tag{3.5}$$

Furthermore, $\mathbf{P}(\tau_y^+ > t)$ is decreasing in t, since

$$\mathbf{P}(\tau_y^+ > t) = \mathbf{P}(_y^+ = t+1) + \mathbf{P}(\tau_y^+ > t+1) \geq \mathbf{P}(\tau_y^+ > t+1). \tag{3.6}$$

Now by (3.6) we see that for $k \geq 0:$

$$r\mathbf{P}_x(\tau_y^+ > kr) \geq \sum_{j=0}^{r-1} \mathbf{P}_x(\tau_y^+ > (k+j)r. \tag{3.7}$$

Hence applying (3.7) and (3.2) to (3.5) gives

$$\mathrm{E}_x[\tau_y^+] = \sum_{t=0}^{\infty} \mathbf{P}_x(\tau_y^+ > t) \leq \sum_{k=0}^{\infty} r\mathbf{P}_x(\tau_y^+ > kr) \leq r\sum_{k=0}^{\infty}(1-\epsilon)^k.$$

Since $0 < \epsilon \leq 1$, this is a converging geometric series. □

Theorem 3.4. *Let X_t be an irreducible Markov chain with transition matrix Q. Then there is a stationary distribution π with:*
$\pi(x) > 0 \; \forall x \in \Omega$ *(1)*
$\pi(x) = \frac{1}{\mathrm{E}_x[\tau_y^+]}$ *(2)*

Proof. Choose some arbitrary but fixed $z \in \Omega$ and define

$$\tilde{\pi} = \mathrm{E}_z(number\,of\,visits\,to\,y\,bef ore\,returning\,to\,z). \tag{3.8}$$

In the formula for this expected value each occurrence of hitting y at a time step less than the return time to z, has value 1. Hence it is simply the sum of the probabilities of these occurrences.

$$\tilde{\pi} = \sum_{t=0}^{\infty} \mathbf{P}_z(X_t = y, \tau_z^+ > t \tag{3.9}$$

The number of possible visits of y before returning to z is bounded by τ_z^+. Therefore $\tilde{\pi} \leq \mathrm{E}_z[\tau_z^+]$, which is finite by the previous Lemma 3.3.

Since the Markov chain is irreducible, for any y there $\exists s, r : Q^s(z,y) > 0$ and $Q^r(y,z) > 0$, which implies $\mathbf{P}_z(X_s = y, \tau_z^+ = s + r) > 0$. Therefore $\tilde{\pi}(y) > 0$.

Now we show that $\tilde{\pi}$ is stationary, which requires to show $\tilde{\pi}(y) = (\tilde{\pi}Q)(y) \forall y \in \Omega$. By the laws of matrix multiplication and the definition of $\tilde{\pi}$ we get

$$(\tilde{\pi}Q)(y) = \sum_{x\in\Omega} \tilde{\pi}(x)Q(x,y) = \sum_{x\in\Omega}\sum_{t=0}^{\infty} \mathbf{P}_z(X_t = x, \tau_z^+ \geq t+1)Q(x,y). \tag{3.10}$$

Due to irreducibility, for any $x \in \Omega$ the inner series

$$\sum_{t=0}^{\infty} \mathbf{P}_z(X_t = x, \tau_z^+ \geq t+1)Q(x,y)$$

converges, because there $\exists N \in \mathbb{N}$ such that $\tau_z^+ = N$. This implies that all summands are zero for $t > N$. Therefore we can switch the order of summation.

Because the event $\{\tau_z^+ \geq t+1\}$ is determined by $X_0, X_1, \ldots, X_t$ it is independent of the event $X_{t+1} = y$, when conditioned on $X_t = x$. Hence we conclude that

$$\mathbf{P}_z(X_t = x, \tau_z^+ \geq t+1)Q(x,y) = \mathbf{P}_z(X_t = x, X_{t+1} = y, \tau_z^+ \geq t+1). \tag{3.11}$$

With these two observations, we can simplify the expression in (3.10) to

$$\begin{aligned}(\tilde{\pi}Q)(y) &= \sum_{t=0}^{\infty}\sum_{x\in\Omega} \mathbf{P}_z(X_t = x, X_{t+1} = y, \tau_z^+ \geq t+1)\\ &= \sum_{t=0}^{\infty} \mathbf{P}_z(X_t = x, \tau_z^+ \geq t+1)\\ &= \sum_{t=1}^{\infty} \mathbf{P}_z(X_t = x, \tau_z^+ \geq t).\end{aligned}$$

Where we use that the events $\{X_t = x\}$ over all $x \in \Omega$ cover the whole space Ω and therefore the inner sum is irrelevant due to basic probability theory. The expression we get is quite similar to the original definition of $\tilde{\pi}(y)$. Put precisely it reads as

$$(\tilde{\pi}Q)(y) = \tilde{\pi}(y) - \mathbf{P}_z(X_0 = y, \tau_z^+ > 0) + \sum_{t=1}^{\infty} \mathbf{P}_z(X_t = y, \tau_z^+ = t)$$
$$= \tilde{\pi}(y) - \mathbf{P}_z(X_0 = y, \tau_z^+ > 0) + \mathbf{P}_z(X_{\tau_z^+} = y). \tag{3.12}$$

The last equality follows, because the return time of z is a uniquely defined integer depending on the random outcome of the Markov chain. Suppose the chain returns to z at time step t', then $\mathbf{P}_z(X_t = y, \tau_z^+ = t) = 0$ for all $t \neq t'$. The last two terms in (3.12) are both

$$\begin{cases} 1, & \text{if } y = z \\ 0, & \text{otherwise} \end{cases}.$$

Thus they cancel each other out in (3.12). In order to make $\tilde{\pi}$ into a stationary distribution we only need to normalize it. Notice that $\sum_{x\in\Omega} \tilde{\pi}(x) = \mathrm{E}_z[\tau_z^+]$ which is greater zero by definition and finite by Lemma 3.3. Thus

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathrm{E}_z[\tau_z^+]}$$

is a *stationary distribution*, with all entries greater than zero, verifying part (1).

Remember that z was chosen arbitrarily and fixed. It´s also valid to choose $z = x$ and run through the whole argument. In this case $\tilde{\pi}(x) = 1$ because the expected number of visits of x before returning to x is exactly 1. Therefore

$$\pi(x) = \frac{1}{\mathrm{E}_x[\tau_x^+]}$$

holds, which finishes the proof of part (2) □

3.3 Uniqueness of stationary distribution

After proving the existence of a stationary distribution for irreducible Markov chains, the next step is to show that it is also unique. In order to prove this, we need a result about harmonic functions.

Definition 3.10. Given a Markov chain with transition matrix Q. A function $h : \Omega \to \mathbb{R}$ is *harmonic* on Ω if

$$h(x) = \sum_{y\in\Omega} Q(x, y)h(y) \quad \forall x$$

.

Remark. A harmonic function h fulfills $Qh = h$.

Lemma 3.5. *Given an irreducible Markov chain X_t with transition matrix Q. If h is harmonic on Ω, then h is a constant function.*

Proof. Without loss of Generality, h attains its maximum at x_0. Suppose that it is strict, meaning

$$h(x_0) > h(x) \; \forall x.$$

Since the Markov chain is irreducible, this implies that $Q(x,x) < 1$ and therefore it exists a $y \in \Omega$ such that $Q(x,y) > 0$. Then it follows that

$$\begin{aligned} h(x_0) &= \sum_{y \in \Omega} Q(x_0,y)h(y) \\ &= Q(x_0,z)h(z) + \sum_{y \neq z} Q(x_0,y)h(y) \\ &< \sum_{y \in \Omega} Q(x_0,y)h(x_0) \\ &< \Big(\sum_{y \in \Omega} Q(x_0,y)\Big)h(x_0) \\ &= h(x_0) \end{aligned}$$

which contradicts $h(x_0) > h(y)$. Hence we conclude $h(x_0) = h(y)$.

Now choose $z \in \Omega$ arbitrary. Then due to irreduciblilty we know that there exists a $s \geq 1$ such that $Q^s(x_0,z) > 0$. This also implies the existence of a path of states $(x_0, x_1, \ldots, x_{s-1}, z)$ with $Q(x_i, x_{i+1}) > 0 \, i = 0, \ldots, s-1$. We can apply the above argument iteratively and conclude that $h(x_0) = h(x_1) = \ldots = h(x_{s-1}) = h(z)$. Hence h is constant. □

Corollary 3.6. *Given an irreducible Markov chain X_t with transition matrix Q. If there exists a stationary distribution π, it is unique.*

Proof. Suppose h is harmonic and therefore constant by Lemma 3.5. The equation $Qh = h$ which is equivalent to $(Q - I)h = 0$, has therefore a kernel with dimension 1. Since the row-rank equals the column-rank of a square matrix we get that the kernel of $vQ = v$ has also dimension 1. Thus all solutions of this equation are spanned by $\{\lambda\pi : \lambda \in \mathbb{R}\}$. The only distribution of this Eigenspace is π, because $\lambda\pi$ is not a distribution for $\lambda \neq 1$. □

The obtained results in this and the last section allow us to explicitly state a formula for the stationary distribution of an irreducible Markov chain.

Corollary 3.7. *An irreducible Markov chain with transition matrix Q has an unique stationary distribution.*

$$\pi(x) = \frac{1}{\mathrm{E}_x[\tau_x^+]} \quad x \in \Omega$$

Proof. This follows immediately by Theorem 3.4 and Corollary 3.6. □

3.4 Convergence theorem

So far, we have shown that any irreducible Markov chain has a unique stationary distribution. In this section, we show that under a further mild constraint a Markov chain will converge for any initial distribution towards its stationary distribution. This additional condition is aperiodicity.

Definition 3.11. For a Markov chain with transition matrix Q let $\mathcal{T}(x) = \{t \geq 1 : Q^t(x,x) > 0\}$ be the set of all time points when it is possible that the chain returns to its starting position x. Then the $\gcd(\mathcal{T}(x))$ is called the *period of x.*

Definition 3.12. A Markov chain is called *aperiodic* if $\gcd(\mathcal{T}(x)) = 1 \quad \forall x \in \Omega$. If it is not aperiodic, it is called periodic.

Lemma 3.8. *If X_t is an irreducible and aperiodic Markov chain with transition matrix Q, then there exists an $r \geq 1$ such that*

$$Q^r(x,y) > 0 \quad \forall x, y \in \Omega$$

Proof. A proof can be found in [8, page 8] or [11, page 195]. □

Before we finally can talk about convergence, we first need to define a measure for the distance between distributions. We look for a well defined metric on the space of distributions.

Definition 3.13. The *total variation* between two distributions μ and ν is defined as

$$\|\mu - \nu\|_{TV} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|$$

.

Proposition 3.9. $\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$

Proof. A proof can be found in [8, page 48]. □

Now we are capable to prove a major result of this thesis, which tells us that an irreducible and aperiodic Markov chain will converge to its stationary distribution at an exponential rate.

Theorem 3.10. *For an irreducible and aperiodic Markov chain with transition matrix Q and stationary distribution π, there exist constants $0 < \alpha < 1$ and $C > 0$ such that*

$$\max_{x \in \Omega} \|Q^t(x, .) - \pi\|_{TV} \leq C\alpha^t \quad \forall t$$

Proof. Let $\Pi = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}$. Then Π is a stochastic matrix since each row is a distribution.

Since the Markov chain is irreducible and aperiodic, by proposition 3.8 there exists some $r > 0$ such that every entry of the matrix Q^r is greater zero. Hence for sufficiently small δ (e.g. $\delta \leq \min \frac{Q^r(x,y)}{\pi(y)}$) it holds that

$$Q^r(x, y) \geq \delta\pi(y) \quad \forall x, y \in \Omega.$$

Now set $\theta = 1 - \delta$ and introduce a stochastic matrix R with

$$\begin{aligned} R &= \frac{Q^r - (1-\theta)\Pi}{\theta} \\ Q^r &= (1-\theta)\Pi + \theta R \end{aligned} \tag{3.13}$$

Since $\pi Q = \pi$, it follows that $\Pi Q^n = \Pi$ for all $n \in \mathbb{N}$. A straight forward computation shows that $R^n\Pi = \Pi$.

Using these observations, we will prove inductively that

$$Q^{rk} = (1-\theta^k)\Pi + \theta^k R^k \quad \forall k \geq 1 \tag{3.14}$$

holds.

For $k = 1$: this holds by equation (3.13)

For the induction step: Assume (3.14) holds for k.

$$\begin{aligned} Q^{r(k+1)} &= Q^{rk}Q^r \\ &= \Big((1-\theta^k)\Pi + \theta^k R^k\Big)Q^r \\ &= (1-\theta^k)\Pi Q^r + \theta^k R^k Q^r \\ &= (1-\theta^k)\Pi + \theta^k R^k\Big((1-\theta)\Pi + \theta R\Big) \\ &= (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}R^{k+1} \\ &= (1-\theta^{k+1})\Pi + \theta^{k+1}R^{k+1} \end{aligned}$$

Where we use (3.13), $\Pi Q^k = \Pi$ and $R^k\Pi = \Pi$ to finish the induction proof.

Now choose $j \in \mathbb{N}$ such that $t = rk + j$ with $0 \leq j < r$ and multiply (3.14) with Q^j. Rearranging terms gives

$$Q^{rk+j} - \Pi = \theta^k (R^k Q^j - \Pi). \tag{3.15}$$

Choose $x \in \Omega$ arbitrary and look at the total variation of $Q^t - \Pi$ of the x-th row.

$$\begin{aligned}
\|Q^t(x,.) - \pi\|_{TV} &= \|Q^t(x,.) - \Pi(x,.)\|_{TV} \\
&= \frac{1}{2} \sum_{y \in \Omega} |(Q^t - \Pi)(x,y)| \\
&= \theta^k \frac{1}{2} \sum_{y \in \Omega} |(R^k Q^j - \Pi)(x,y)| \\
&= \theta^k \|R^k Q^j(x,.) - \Pi(x,.)\|_{TV} \\
&\leq \theta^k
\end{aligned}$$

Here we have used the specific construction of $Q^t - \Pi$ 3.15, Proposition 3.9 twice and the fact that the total variation is bounded by 1.

We finish the proof with defining the constants $\alpha = \sqrt[r]{\theta}$ and $C = \alpha^{-r}$ and notice that $\theta^k \leq C\alpha^t$. □

This result is great, because it provides that π is also a limiting distribution.

Corollary 3.11. *Given an irreducible and aperiodic Markov chain with transition matrix Q. The stationary distribution π is a limiting distribution.*

Proof. Take any probability distribution p^0 on Ω. Then it holds

$$\lim_{t \to \infty} p^0 Q^t = p^0 \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \pi.$$

□

3.5 General state space

In Section 4 we will introduce the Metropolis Hastings algorithm which uses Markov chain theory to approximate distributions. In most applications these distributions are continuous. Therefore we need Markov chain theory on Euclidean spaces. So far we had the restriction that the state space Ω must be finite. For that special case we could derive the wonderful result, that every irreducible, aperiodic Markov chain has a unique stationary distribution towards which it converges at an exponential rate. Fortunately the extension from Ω finite to $\Omega = \mathbb{R}^d$ is more of a technical nature and all the results we derived so far also hold for these state spaces. It is actually possible to extend the theory to all spaces which have a finitely generated sigma algebra.

A brief and intuitive introduction into Markov chain theory on general state spaces, can be found in [4, Chapter 4] and for a detailed elaboration take a look into [10] or [5].

An essential difference is that Ω is now a probability space in the sense of measure theory and Q is no longer a transition matrix, but now a transition kernel which means that

$$\begin{aligned} &Q(x, A) = \mathbf{P}(X_{t+1} \in A | X_t = x) \\ &\forall x \in \mathbb{R}^d, A \subset \mathbb{R}^d, \quad t \geq 0 \end{aligned}$$

.

The main takeaway is that the gained results still hold. More details can be found in the above mentioned literature.

3.6 Reversibility

In the final subsection of the Markov chain theory part we briefly introduce reversible distributions and show that any reversible distribution is actually stationary. That result is the last building block of Markov chain theory, that one needs to introduce the Metropolis Hastings algorithm.

Definition 3.14. Given a Markov chain with transition matrix Q with state space Ω. If a distribution π fulfills:

$$\pi(x)Q(x, y) = \pi(y)Q(y, x) \quad \forall x, y \in \Omega \tag{3.16}$$

then π is called *reversible*. The equation (3.16) is called the detailed balance equation.

Proposition 3.12. *Any reversible π is stationary.*

Proof. Take $x \in \Omega$ arbitrary. Then

$$(\pi Q)(x) = \sum_{y \in \Omega} \pi(y)Q(y, x) = \sum_{y \in \Omega} \pi(x)Q(x, y) = \pi(x) \sum_{y \in \Omega} Q(x, y) = \pi(x)$$

holds. Since x was arbitrary, this holds for all. □

4 The Metropolis-Hastings algorithm

4.1 Introduction

The Metropolis-Hastings algorithm is based on the work of Metropolis (1953) [9] and Hastings (1970) [6] and belongs to the class of Markov Chain Monte Carlo (MCMC) algorithms. As the name points out, this algorithm combines Markov chain theory and Monte Carlo sampling.

The main purpose of the Metropolis Hastings algorithm is to draw samples from a distribution from which direct sampling is not possible. A sample is drawn in such a way that the sample approximates the distribution closely for a sufficiently large sample size. The distribution from which we want to sample is called target distribution and in most applications it is continuous. A field where such problems do naturally occur is Bayesian statistics. The object of fundamental interest in Bayesian statistics is the so called *posterior distribution* from which direct sampling is hardly ever possible. Hence the Metropolis Hastings algorithm is a widely used tool in this field.

The core idea of the algorithm is to turn the convergence theorem for irreducible and aperiodic Markov chains upside down. Instead of finding the stationary distribution of a Markov chain, it constructs a Markov chain which is irreducible, aperiodic and has the target distribution as its limiting distribution.

In section 4.2 we will introduce the algorithm in detail and section 4.3 gives a proof that the chain generated by the algorithm converges towards the target distribution. The last section 4.4 first shows approximations of mixed normal distributions and finishes with sampling the posterior distribution in a Bayesian regression model.

4.2 The algorithm

The concept and ideas presented in this and the next chapter are based on [4] and [1]. As in the introduction already set out, the purpose of the Metropolis Hastings algorithm is to get a good approximation of some distribution π (called target distribution) over some space Ω (Euclidean space). Therefor the algorithm constructs an irreducible, aperiodic Markov chain for which the target distribution π is reversible. This task is accomplished in the following way: The state in period 0 is chosen randomly. For each new time step t, one draws a candidate state x^{cand} from a conditional distribution $q(.|X_{t-1})$. This distribution is called the proposal distribution. The only restriction on $q()$ is that

$$q(x|y) > 0 \quad \forall\, x, y\, \in \Omega. \tag{4.1}$$

Given a candidate x^{cand}, the algorithm then has to decide if it should switch from the current state to the proposed one. This decision is made by looking at the following acceptance probability.

$$\alpha(x_{t-1}, x^{cand}) = \min\Big(1, \frac{\pi(x^{cand})q(x_{t-1}|x^{cand})}{\pi(x_{t-1}q(x^{cand}|x_{t-1})}\Big) \tag{4.2}$$

The acceptance probability compares the probability of being in the current state versus being in the proposed state with respect to the target and proposal distribution. With probability α the proposed state x^{cand} is accepted and with $1 - \alpha$ the chain stays in the current state. The algorithm is doing that by comparing α to a randomly drawn number u from the Uniform distribution of the unit interval. If $u \leq \alpha$ the proposed state is accepted. Intuitively speaking, the constructed chain shall spend more time in regions of Ω where π has higher probability. This procedure is repeated N-times.

Written in algorithmic notation it simply reads as:

```
Choose $x_0 \in \Omega$
for $t \leftarrow 1, N \in \mathbb{N}$ do
    Draw a proposal $x^{cand}$ from $q(.|x_{t-1})$
    Draw $u$ from $\mathcal{U}(0,1)$
    if $u \leq \alpha$ then
        $x_t \leftarrow x^{cand}$
    else
        $x_t = x_{t-1}$
    end if
end for
```

In the special case where the proposal distribution is symmetric, meaning $q(x_{t-1}|x^{cand}) = q(x^{cand}|x_{t-1})$, the algorithm is the original Metropolis algorithm.

4.3 Convergence and Implementation

Theorem 4.1. *The Metropolis Hastings algorithm generates a Markov chain which has the target distribution π as its limiting distribution.*

Proof. The chain is irreducible by the restriction on the proposal distribution (4.1). For any x, y in Ω it is possible to get from state x to state y in 1 time step by construction.

Also by construction it is aperiodic, since for any state x it is possible to stay at that state. Therefor all periods are 1.

Now the convergence theorem tells us, that the unique stationary distribution of the generated chain is also its limiting distribution. The remaining thing to show is that the target distribution is actually the stationary distribution of the Metropolis Hastings chain. We proof that by showing that the target distribution π is reversible, because then Proposition 3.12 gives us the desired result.

The transition kernel of the Metropolis Hastings chain $Q(x, y)$ for moving from x to y equals $q(y|x)\alpha(x, y)$. The following straight forward computation shows

$$\begin{aligned}
\pi(x)Q(x,y) &= \pi(x)q(y|x)\alpha(x,y) \\
&= \pi(x)q(y|x)\min\left(1, \frac{\pi(y)q(x|y}{\pi(x)q(y|x}\right) \\
&= \min(\pi(x)q(y|x), \pi(y)q(x|y)) \\
&= \pi(y)q(x|y)\min\left(\frac{\pi(x)q(y|x)}{\pi(y)q(x|y)}, 1\right) \\
&= \pi(y)q(x|y)\alpha(y,x) \\
&= \pi(y)Q(y,x)
\end{aligned}$$

that π is reversible. □

This result guarantees convergence of the algorithm for $t \to \infty$, but when it comes to implementing the algorithm one has only finitely many time steps N to gain sufficient convergence. Hence it is always necessary to carefully analyze the output and use some test procedures to gain confidence that the algorithm did converge. Some standard tricks are to use a burn-in phase, meaning to throw away the first n runs of the algorithm to account for the fact that the initial state does effect at which states the chain will be in the beginning. Another one is to run the algorithm several times with different initial values and compare the outputs. Also very important is to check how different proposal distributions influence the output and to have a look at the acceptance rate (share of accepted candidate states). It exists an extensive literature about implementation issues of the Metropolis Hastings algorithm, e.g. [4].

4.4 Applications

The next section shows approximations of univariate mixed normal distributions. These are kind of artificial examples, because we use the Metropolis Hastings algorithm to sample from distributions for which we actually know the exact solutions. However in this setting we can properly display the approximation behaviour of the Metropolis Hastings algorithm since we know exactly how the target distribution looks like. In the last part of the thesis we will present an application of the algorithm in a simple linear Bayesian regression model. These examples are implemented in R and the Code is provided in [12].

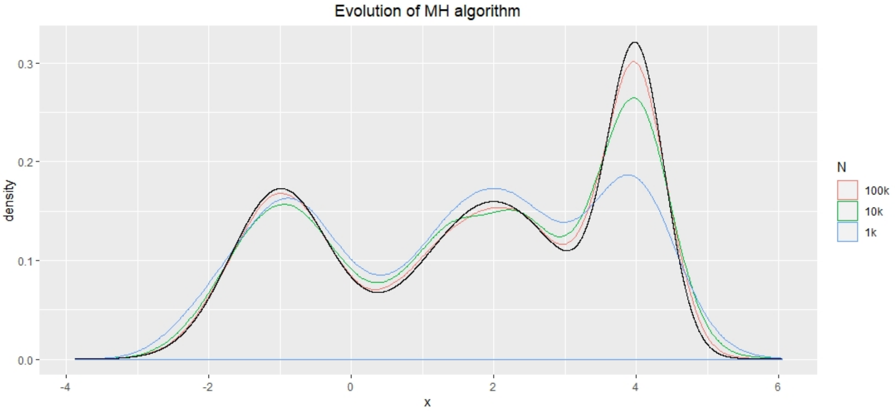
4.4.1 Approximate mixed normal distributions

As already mentioned above, this chapter will present approximations of two mixed normal distributions with the Metropolis Hastings algorithm. These are show case examples to illustrate the approximation behaviour of the algorithm.

The first example is the approximation of the following mixed normal distribution:

$$\pi \sim 0.3\mathcal{N}(-1, 0.7) + 0.4\mathcal{N}(2, 1) + 0.3\mathcal{N}(4, 0.4)$$

Although the choice of the proposal distribution is from a mathematical standpoint not crucial, because due to the convergence theorem eventually the Metropolis chain will converge towards the target distribution. However the choice is very important from a computational perspective, because we need to get sufficient convergence within N steps. Therefore the proposal distribution was set to $q(x^{cand}|x_{t-1} = \mathcal{N}(x_{t-1}, 3)$, because it gives an acceptance rate of slightly above 50% and therefore the mixing of the chain over the support is sufficient. The algorithm was initialised with $x_0 = -10$ and ran $N = 100k$ steps. In order to get rid of the effects of the randomly chosen initialization point, the burn-in phase was set to $1k$.

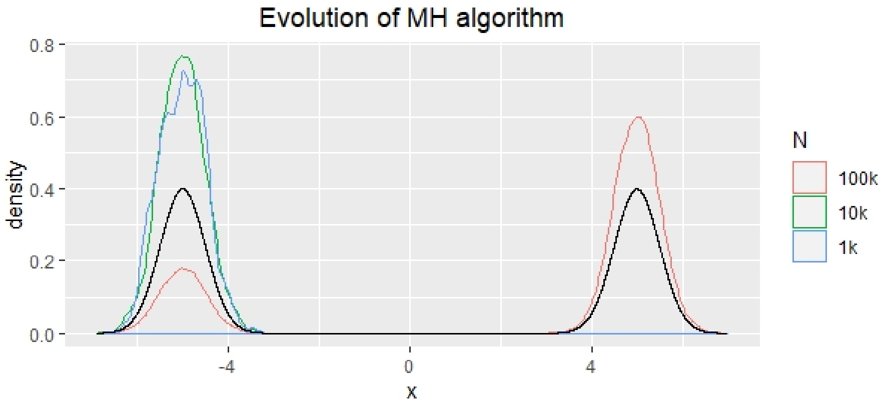


After $1k$ steps the approximation is at the peak at $x = 4$ far away from the target distribution, but already quite okay at most parts of the support. One can see that as the algorithm goes on it gets astonishingly close to the target distribution. Hence in this case the algorithm performs well.

In the next example the target distribution has "holes" in the support. It looks as follows:

$$\pi \sim \frac{1}{2}\mathcal{N}(-5, 0.5) + \frac{1}{2}\mathcal{N}(5, 0.5)$$

Running the algorithm with the same specifications as above gives the following result.



In this case the algorithm tends to get stuck at one of the two regions where the density is high. Therefore the algorithm has a hard time to converge.

4.4.2 Bayesian linear regression model

In order to show how the Metropolis Hastings algorithm is applied in Bayesian statistics we first give a brief introduction to Bayesian linear regression. Therefore we start by laying out the mainstream approach about linear regression, called frequentist statistics. Then we illustrate the Bayesian perspective on linear regression. The part about frequentist statistics is based on [2] and the introduction to Bayesian statistics is based on [7]. After the introduction we conclude with showing a simple linear Bayesian regression model for simulated data.

Definition 4.1. Let $X \in \mathbb{R}^{n \times k}$, with linearly independent columns. Take fixed $\beta_0 \in \mathbb{R}^k, \sigma_0 > 0$. Then

$$X\beta_0 + \epsilon, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_0), i = 1, \ldots, n$$

defines a *data generating process.* The fixed β_0 and σ_0 are referred to as the true parameters of the data generating process. The columns of X are called explanatory or independent variables.

Suppose one has given some data (y, X) where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times k}$ with linearly independent columns. For simplicity let us assume that the explanatory variables of X do actually represent the data generating process which generated y. Then we set up the following model

$$Model: \quad y = X\beta + u$$

because we assume that y was generated by X. In reality the choice of the model is actually a crucial part. Then the task at hand is to find good estimates $\hat{\beta}$ of the model parameter β. The most commonly used estimation method is Ordinary Least Squares which says

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

The Ordinary Least Squares estimator fulfills $\mathbf{E}[\hat{\beta}_{OLS}] = \beta_0$ and $\lim_{n\to\infty} \hat{\beta}_{OLS} = \beta_0 \, in \, probability$. However notice that $\hat{\beta}$ is not directly treated as a random variable, but the observed data is seen as being randomly observed. Hence the expected value and the probability limit are taken over the space of data samples. In the given Definition of a data generating process and the way it is used in frequentist statistics, the parameters β_0, σ_0 are not random, but the observed data is treated as random.

In contrast to that is Bayesian statistics, which has a fundamentally different view about how the data generating process and modelling it is connected. In a Bayesian view, the parameters $\theta = (\beta, \sigma)$ are thought of as random variables and the observed data is seen as not random. This perspective is justified in the following two ways:

The problem of estimating a linear regression model invokes a lot uncertainty about the parameters (also about the model choice). This uncertainty can be formalized with probability theory, by relying on the subjective interpretation of probability. The only known thing is the observed data. Hence what we are looking for is the conditional probability $p(\theta|y)$. This argumentation is from [7, p.2].

Another line of argumentation points out that the parameters of the data generating process must be actively chosen in order to be non random. Hence the logical follow up question is "Who did choose them and how?". It can't be chosen by some evolutionary process, because such processes are random by their nature. A common way of answering this question is then "Only God knows the true parameters" (a phrase that regularly appears in statistic lectures), implying that some non-human power actively set the true parameters of the data generating process. This explanation is inconsistent, since it depends on the existence of "God". It is more plausible to think that the parameters are determined randomly (e.g. by Evolution, History, etc.) Or putting it differently "If God exists, she might have also rolled the dice".

After accepting the Bayesian perspective that the parameters are random variables and only the data is known, the path to take is quite straight forward. Of fundamental interest is the conditional probability $p(\theta|y)$, which is called the posterior. Applying Bayes theorem gives:

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} p(\theta) \tag{4.3}$$

With the posterior it is possible to answer in a full probability framework what the expected value and the quantiles of the parameters are. This and many more tasks of inference can be done with the posterior.

We call $p(y|\theta)$ the likelihood and $p(\theta)$ is called the prior. If one ignores $p(y)$ in (4.3) (which is usually done, because it is not necessary for the analysis of the posterior) one can see that the posterior is proportional to the likelihood times the prior.

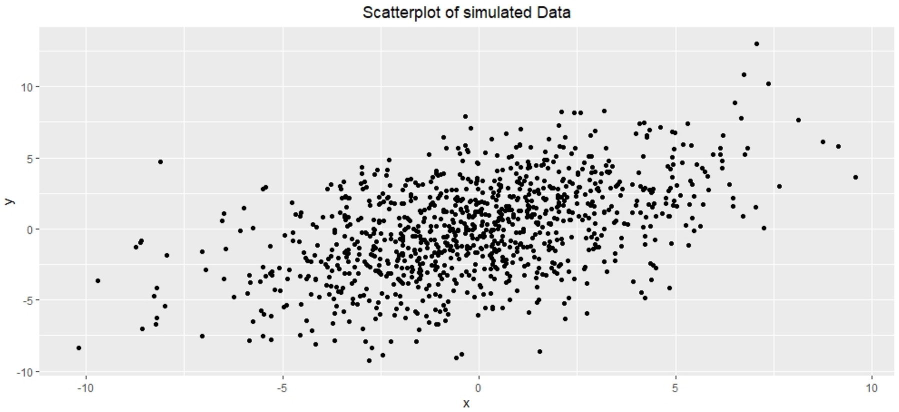
$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{4.4}$$

According to this structure it is hardly ever possible to solve the posterior analytically. Hence we need a suitable tool to approximate it and a commonly used one is the Metropolis Hastings algorithm.

The following example shall demonstrate how the Metropolis Hastings algorithm is applied in Bayesian regression models. First let us simulate the following data.

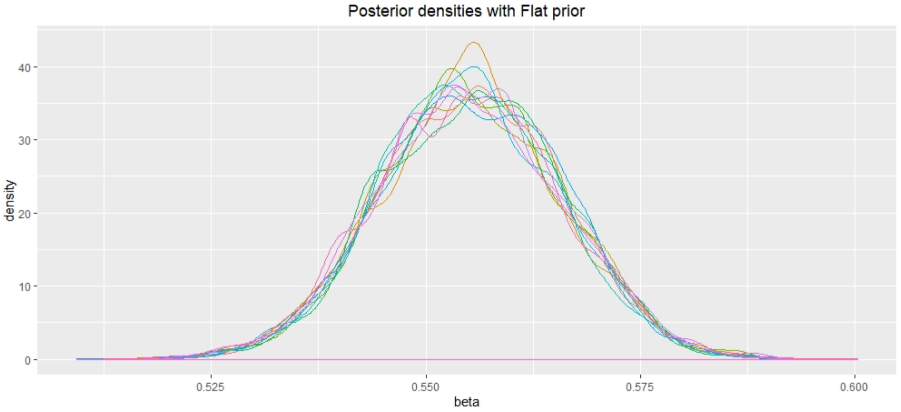
$$\begin{aligned} &x \sim \mathcal{N}(0, 3) \\ &y = x\frac{1}{2} + \epsilon, \quad \epsilon_i \sim \mathcal{N}(0, 3), i = 1, \ldots, n \end{aligned}$$

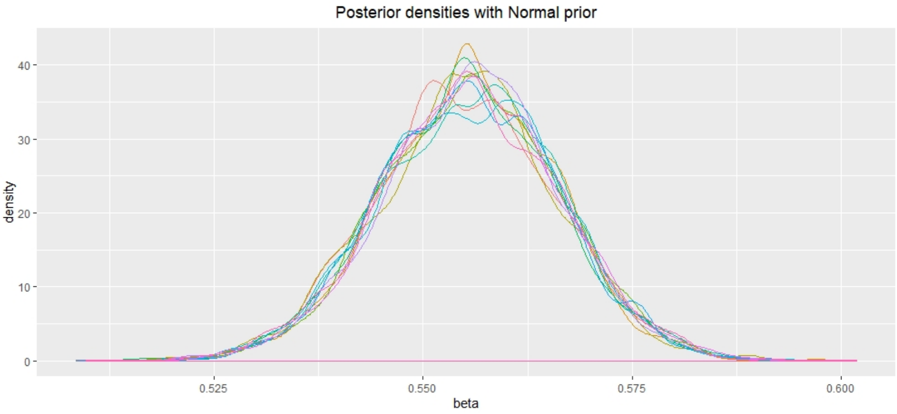
Where we set the linear coefficient $\beta_0 = \frac{1}{2}$, the number of observations to $n = 1000$ and the error variance $\sigma_0 = 3$. The following plot shows the simulated data.



In order to see how robust the posterior is concerning the choice of the prior, we compare the results for a flat prior $p(\beta) = \mathcal{U}(-30, 30)$ and a normal prior $p(\beta) = \mathcal{N}(3, 1)$.

We use the Metropolis Hastings algorithm to sample from the posterior of β for both priors. The chosen proposal distribution is $\mathcal{N}(x_{t-1}, 0.05)$, which was chosen by try and error in order to get an appropriate acceptance rate of around 25%. The second example of Section 4.4.1 showed that the Metropolis Hastings algorithm has a hard time to converge if the target distribution has disconnected areas of positive support. Therefore the implementation includes the following two strategies. The algorithm is run several times with initialisation points for β being $-25, -20, \ldots, 20, 25$ and in each run of length $N = 12k$ the first $2k$ states are thrown away (Burn-in phase). This procedure seems suitable to detect if the posterior distribution has multiple disconnected "hot spots". The following figures show the results:





For both priors the results are very similar. Namely, that no matter which initialisation point is chosen, the posteriors are very close to each other. Hence we are very certain that there is no other area where the posterior has positive probability. As the Bayesian estimate of β we take the mean expected value of all the posteriors, which is 0.55536 with the flat prior and 0.55547 with the normal prior. Essentially the estimates are identical and also the frequentist OLS estimate is with 0.55529 more or less the same. Our Bayesian estimate is very close to the true $\beta_0 = \frac{1}{2}$.

References

[1] S Chib and E. Greenberg. “Understanding the Metropolis-Hastings Algorithm”. In: *The American Statistician* 49 (1995), pp. 327–335.

[2] Russell Davidson, James G MacKinnon, et al. *Econometric theory and methods.* Vol. 5. Oxford University Press New York, 2004.

[3] A Freedman. “Convergence theorem for finite Markov chains”. In: (2017).

[4] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1996.

[5] M. Hairer. “Convergence of Markov Processes”. In: (2016).

[6] W. K. Hasting. “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57 (1970), pp. 97–109.

[7] G. Koop. *Bayesian Econometrics.* John Wiley & Sons Ltd., 2003.

[8] D. Levin, Y. Peres, and E. Wilmer. *Markov chains and mixing times.* 2017.

[9] N. Metropolis et al. “Equations of state calculations by fast computing machine”. In: *J. Chem. Phys.* 21.6 (1953), pp. 1087–1091.

[10] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Cambridge University Press, 2009.

[11] T. Müller-Gronbach, E. Novak, and K. Ritter. *Monte Carlo Algorithmen.* Springer, 2012.

[12] L. Peham. *RCode: Markov-chain-theory-and-the-Metropolis-Hastings-algorithm.* URL: `https://github.com/lpeham93/RCode-Markov-chain-theory-and-the-Metropolis-Hastings-algorithm`. (accessed: 10.02.2020).