Avd. Matematisk statistik

KTH Teknikvetenskap

Sf2955 COMPUTER INTENSIVE METHODS
MCMC
On the Metropolis-Hastings Algorithm on Continuous State Spaces
2009
Timo Koski

1 Introduction

1.1 MCMC

MCMC is a powerful method for exploring (by simulation) a high-dimensional distribution. Problems arising in Bayesian inference and other fields frequently lead to highdimensional distribution that are (absolutely) continuous, i.e., have a density. Hastings (1970) suggested the use of the discrete Metropolis-Hastings algorithm by discretization of the pertinent densities, but we do not follow his lead in this lecture.

In fact it seems to be natural to base the definítion and theoretical analysis of MCMC on the theory of Markov chains on a general space. In this lecture we give a brief and informal summary of general Markov chains as expounded in (Nummelin 1984, Meyn & Tweedie 1993). Our summary follows the outline in (Robert and Castella 1999, Tierney 1994).

The approach to Markov chain theory in discrete time and on a general (continuous) state space is to start with a transition kernel, which gives us $P(A|x)$, the conditional probability distribution of moving from x to a set A.

A major concern in the theory is to find conditions under which there exists an invariant distribution for the chain, and conditions which allow iterations

of the transition kernel to converge to this invariant distribution. This is an important concern for chains defined by the MCMC algorithms, too.

2 Markov Chains with Continuous State Space

2.1 Transition kernel

First we introduce some notation and facts about transition kernels of Markov chains with a general (uncountable) state space S and discrete time. The state space can be thought of as $\mathcal{R}^n$, the n-dimensional Euclidean space. (In this case $x \in S \Leftrightarrow x = (x_1, \ldots, x_n)$.)

2.2 Notations & Facts

Definition 2.1 *[***Transition kernel***]* A *transition kernel* is a function K defined on $S \times \mathcal{B}(S)$ ($\mathcal{B}(S)$ = a countably sigma-algebra of subsets of S) with values in $[0, 1]$ such that

(i) $\forall x \; K(x, \cdot)$ is a probability measure.

(ii) $\forall A \in \mathcal{B}(S)$, $K(x, A)$ is measurable.

■

Thus $K(x, \cdot)$ is a version of the conditional distribution

$$P(X_{n+1} \in \cdot \mid X_n = x)$$

of X_{n+1} given X_n. This formulation permits the probability of transition to be a singleton, a set of the form $\{y\}$, to be positive, i.e.,

$$K(x, \{y\}) > 0$$

We compound the notations by letting the $K(\cdot, \cdot)$ to denote also the transition density $K\left(x, x^{'}\right)$ in the sense that

$$P(X_{n+1} \in A | X_n = x) = K(x, A) = \int_A K\left(x, x^{'}\right) dx^{'} \tag{2.1}$$

holds for all $A \in \mathcal{B}(S)$, when the density exists.

Definition 2.2 *[Markov chain]* Given a transition kernel K, a sequence $\{X_n\}_{n\geq 0}$ is a *Markov chain*, if for any n and any $x_0, \ldots, x_n$, the conditional distribution of X_{n+1} given $X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0$ is

$$P\left(X_{n+1} \in A \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0\right) = P\left(X_{n+1} \in A \mid X_n = x_n\right)$$
$$= K\left(x_n, A\right) = \int_A K\left(x_n, x\right) dx.$$

■

The chain is a time-homogeneous, if the transition density does not depend on n.

Example 2.1 (Random Walk) Let

$$X_{n+1} = X_n + Z_n, \tag{2.2}$$

when Z_n are I.I.D.,and independent of X_0. Then X_{n+1} is conditionally independent of $X_{n-1}, \ldots, X_0$ given X_n, and the ensuing Markov chain $\{X_n\}_{n\geq 0}$ is called a **random walk**. Assume that Z_n has the density $g(x)$. Then

$$P\left(X_{n+1} \leq y \mid X_n = x\right) = P\left(x + Z_n \leq y \mid X_n = x\right)$$
$$= P\left(Z_n \leq y - x\right) = \int_{-\infty}^{y-x} g(z)dz.$$

Thus the transition density is

$$K(x, y) = \frac{\partial}{\partial y} P\left(X_{n+1} \leq y \mid X_n = x\right) = g(y - x). \tag{2.3}$$

■

Example 2.2 *[Gaussian Autoregressive Process of Order 1, AR(1)]* Let $\theta \in \mathcal{R}$, and let

$$X_{n+1} = \theta X_n + Z_n, \tag{2.4}$$

where $Z_n \sim N\left(0, \sigma^2\right)$, and I.I.D., and independent of X_0. Then X_{n+1} is indeed Gaussian and conditionally independent of $X_{n-1}, \ldots, X_0$ given X_n. The transition probability density is obtained as in (2.3), and is

$$K\left(x, y\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2}}. \tag{2.5}$$

■

2.3 Iterates of the transition kernel

We have also

$$P\left(X_1 \in A_1 | X_0 = x\right) = \int_{A_1} K\left(x, x^{'}\right) dx^{'} = K\left(x, A_1\right)$$

$$P\left(\left(X_1, X_2\right) \in A_1 \times A_s | X_0 = x\right) = \int_{A_1} K\left(y, A_2\right) K\left(x, y\right) dy$$

$$\vdots$$

$$P\left(\left(X_1, X_2, \ldots, X_n\right) \in A_1 \times \ldots \times A_n | X_0 = x\right)$$

$$= \int_{A_1} \int_{A_2} \ldots \int_{A_{n-1}} K\left(x, y_1\right) \ldots K\left(y_{n-2}, y_{n-1}\right) K\left(y_{n-1}, A_n\right) dy_1 \ldots dy_{n-1}.$$

If we set $K^1\left(x, A\right) = K\left(x, A\right)$, the kernel for n transitions is given by

$$P\left(X_n \in A | X_0 = x\right) = K^n\left(x, A\right) = \int_S K^{n-1}\left(y, A\right) K\left(x, y\right) dy.$$

We get with this notation the **Chapman-Kolmogorov equation**

Lemma 2.3 For all $n, m \in N \times N$, $x \in S$, $A \in \mathcal{B}(S)$,

$$P\left(X_{n+m} \in A \mid X_0 = x\right) = K^{m+n}\left(x, A\right) = \int_S K^n\left(y, A\right) K^m\left(x, y\right) dy.$$

■

2.4 Stopping times

One of the most important tools of probability calculus is the notion of a stopping time.

Definition 2.3 *[***Stopping time***]* Take $A \in \mathcal{B}(S)$, and

$$\tau_A = \min\{n \geq 1 \mid X_n \in A\}. \tag{2.6}$$

and

$$\tau_A = +\infty,$$

if $X_n \notin A$ for any n.

■

In words, τ_A is the first time the chain enters A, and is called the *stopping time* at A. We define also

$$\eta_A = \sum_{n=1}^{\infty} I_A(X_n), \tag{2.7}$$

which is the number of visits of the chain in A.

We will be invoking the quantitities

$$E[\eta_A],$$

or the average number of visits in A, and

$$P(\tau_A < \infty),$$

which is the probability of return to A in a finite number of steps.

2.5 Classification of states

2.5.1 Irreducibility

The chain is called **irreducible**, if all states communicate,

$$\forall x, y \in S \times S, P(\tau_y < +\infty \mid X_0 = x) > 0.$$

where τ_y, the first time y is visited, when the chain starts in x, is defined as in (2.6).

This is not a good definition of irreducibility, and has to be corrected by the following.

Definition 2.4 *[ϕ-**irreducibility**]* Given a measure ϕ, the Markov chain $\{X_n\}_{n\geq 0}$ with transition kernel $K(x,y)$ is ϕ-**irreducible**, if for every $A \in \mathcal{B}(S)$ with $\phi(A) > 0$ there exists an $n = n(x,A)$ such that

$$K^n(x,A) > 0 \text{ for all } x \in S$$

$$\Leftrightarrow P(\tau_A < \infty | X_0 = x) > 0.$$

■

The literature cited invokes at this stage the existence of a maximal ϕ -measure, and formulates much of the results in terms of the maximal irreducibilty measure, but we refrain from doing so here.

2.5.2 Transience and Recurrence

From an algorithmic point of view a Markov chain must have stability properties. Irreducibility ensures that every set A will be visited by the MC $\{X_n\}_{n\geq 0}$, but this property is too weak to ensure that the sample path of $\{X_n\}_{n\geq 0}$ will enter A often enough. The literature cited introduces here also the Harris recurrence, which we skip here.

Definition 2.5 *[Recurrence]* A Markov chain $\{X_n\}_{n\geq 0}$ is **recurrent**, if

- there is a measure ϕ such that $\{X_n\}_{n\geq 0}$ is ϕ-irreducible, and
- for every $A \in \mathcal{B}(S)$ such that $\phi(A) > 0$ and

$$E[\eta_A] = +\infty \text{ for every } x \in A,$$

where η_A is defined in (2.7)

■

Definition 2.6 *[Transience]* A Markov chain $\{X_n\}_{n\geq 0}$ is **transient**, if

- there is a measure ϕ such that $\{X_n\}_{n\geq 0}$ is ϕ-irreducible, and
- there exist $A_i \in \mathcal{B}(S)$ such that $\phi(A_i) > 0$ and

$$E[\eta_{A_i}] < M_i \text{ for every } x \in A,$$

where η_{A_i} is defined in (2.7), and

$$S = \bigcup_i A_i.$$

■

We have the following theorem.

Proposition 2.4 A ϕ-irreducible chain is either recurrent or transient.

■

2.6 Invariant measure

Definition 2.7 *[Invariant measure]* A σ-finite measure is **invariant** for the transition kernel $K(x, y)$, and for the associated Markov chain, if

$$\pi(B) = \int_S K(x, B)\, \pi(dx) \quad \forall B \in \mathcal{B}(S). \tag{2.8}$$

■

If the density (again denoted by π) exists, the definition of an invariant measure can be stated as

$$\pi(y) = \int_S K(x, y)\, \pi(x)\, dx. \tag{2.9}$$

When there exists an *invariant* probability *measure* for a ϕ-irreducible chain, the chain is called **positive**. Recurrent chains that do not allow for a finite measure are called **null recurrent**.

Example 2.5 *[Invariant density for AR(1)]* Assume that $\mid \theta \mid < 1$ in

$$X_{n+1} = \theta X_n + Z_n, \tag{2.10}$$

where $Z_n \sim N(0, \sigma^2)$, and I.I.D., and independent of X_0. In example 2.2 we found that

$$K(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2}}. \tag{2.11}$$

Thus (2.9) becomes

$$\pi(y) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2}} \pi(x)\, dx. \tag{2.12}$$

It can be checked that the invariant distribution, the density of which satisfies (2.12), is $N\left(0, \frac{\sigma^2}{1-\theta^2}\right)$.

■

If $X_0 \sim \pi$, then $X_n \sim \pi$ for all n, and if π is finite measure, then the chain is called stationary.

Proposition 2.6 If the chain $\{X_n\}_{n \geq 0}$ is positive, then it is recurrent.

■

2.7 Reversible Markov Chains

$\{X_n\}_{n\geq 0}$ is an MC with the kernel $K(x,y)$.

Definition 2.8 If for any A and B in $\mathcal{B}(S)$ and any $n \geq 0$

$$P\left(X_n \in A, X_{n+1} \in B\right) = P\left(X_n \in B, X_{n+1} \in A\right), \tag{2.13}$$

then we say that the Markov chain $\{X_n\}_{n\geq 0}$ is **reversible**.

■

It turns out that this is equivalent to the **detailed balance condition**, i.e., there is a function f satisfying

$$K\left(y,x\right) f(y) = f(x) K\left(x,y\right) \tag{2.14}$$

for every $(x,y) \in S \times S$.

Proposition 2.7 Suppose that a Markov chain with the transition kernel $K(x,y)$ satisfies the detailed balance condition (2.14) with f, which is a probability density function. Then

1) The density f is the invariant density of the chain.

2) The chain is reversible.

Proof:

1)

$$\begin{aligned}\int_S K\left(y,B\right) f(y)dy &= \int_S \int_B K\left(y,x\right) f(y)dxdy \\ &= \int_S \int_B K\left(x,y\right) f(x)dxdy \\ &= \int_B f(x) \int_S K\left(x,y\right) dydx = \int_B f(x)dx,\end{aligned}$$

since $\int_S K\left(x,y\right) dy = 1$.

2) To prove the reversibility from the detailed balance condition (2.14) we note that

$$P\left(X_n \in A, X_{n+1} \in B\right) = \int_A f(x) P\left(X_{n+1} \in B | X_n = x\right) dx$$

$$= \int_A \int_B f(x) K\left(x, y\right) dy dx$$

$$= \int_A \int_B f(y) K\left(y, x\right) dy dx$$

$$= \int_B \int_A f(y) K\left(y, x\right) dx dy = P\left(X_n \in B, X_{n+1} \in A\right),$$

where we used also Fubini's theorem.

■

In fact (2.14) and (2.13) are equivalent, when densities exist; from (2.13) we obtain (2.14) by selecting A and B suitably and by differentiating.

3 MCMC for Continuous Spaces

3.1 Introduction

The MCMC theory turns the preceding around: the invariant distribution is known, perhaps up to a constant multiple, it is the target density from which samples are desired, and the problem is find the desired transition kernel. But we should be able to check that the Markov chain obtained is a positive chain, as we may want to compute various integrals by Monte Carlo.

3.2 The Metropolis Problem Again

Let f be an arbitrary probability density function, **target density**, on S, i.e.,

- $f(x) \geq 0$ for all $x \in S$.
- $\int_S f(x) dx = 1$.

The *Metropolis problem* is to give a Markov chain $\{X_n\}_{n \geq 0}$, i.e., the associated kernel, such that f is its invariant density (distribution).

3.3 Metropolis-Hastings Algorithm

Let

$$q\left(y|x\right)$$

be a conditional density function, i.e.,

- $q(y|x) \geq 0$ for all $x, y \in S \times S$.
- $\int_S q(y|x)dy = 1$.

The various factors influencing selection of $q\left(y|x\right)$ are discussed below in section 5.

Definition 3.1 *[Metropolis-Hastings Algorithm]* $q\left(y|x\right)$ is a conditional probability density. Given that $X_n = x_n$

1. Generate $Y_{n+1} \sim q\left(y|x_n\right)$.
2. Take

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{with probability } \rho\left(x_n, Y_{n+1}\right) \\ i & \text{with probability } 1 - \rho\left(x_n, Y_{n+1}\right), \end{cases}$$

   where

$$\rho(x, y) = \min\left\{1, \frac{f(y)q(x|y)}{f(x)q(y|x)}\right\}. \tag{3.1}$$

3. $X_{n+1} \mapsto x_n$ and return to 1.

The distribution q is called the *proposal* distribution.

■

3.3.1 Comments on the Metropolis-Hastings Algorithm

- The calculation of $\rho(x, y)$ does not require the knowledge of the normalizing constant in the target distribution.
- in case $q(y|x)$ is symmetric (Metropolis Algorithm), the probability of acceptance depends only on the ratio $f(y)/f(x)$. If $f(y) > f(x)$, the chain moves to y; otherwise it moves to with the probability given by $f(y)/f(x)$. Hence, if the jump goes 'uphill', it is always accepted, if it goes downhill, it is accepted with a non-zero probability.

3.3.2 Properties of the Metropolis-Hastings Algorithm

Let us next add a technical assumption, the formulation of which will require a definition. Let $\mathcal{E}_f$ = the **support** of f, the $\mathcal{E}_f$ smallest closed set that contains all x such that $f(x) > 0$,

$$\mathcal{E}_f = \text{cl}\{x \in S \mid f(x) > 0\}.$$

We are going to assume that $\mathcal{E}_f$ is *connected*. Connectedness means that for any two points x and y in $\mathcal{E}$ there is a path connecting x and y that is included inside $\mathcal{E}$.

Proposition 3.1 Let $q(y|x)$ be any conditional distribution, the support of which, $\mathcal{E}_q$, contains the support $\mathcal{E}_f$ of f, where $\mathcal{E}_f$ is assumed to be connected. Then the Markov chain $\{X_n\}_{n\geq 0}$ produced by the Metropolis-Hastings algorithm in definition 3.1 above has the density f as the density of the invariant distribution.

Proof: The proof is given in Appendix A. ∎

Let

$$\phi(A) = \int_A f(x)dx \tag{3.2}$$

We say that an MC is f - *irreducible*, if it is ϕ-reducible for the measure ϕ defined in (3.2).

Proposition 3.2 Suppose that the Markov chain $\{X_n\}_{n\geq 0}$ produced by the Metropolis-Hastings algorithm in definition 3.1 is f-irreducible. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{n=1}^{n} h(X_n) = \int_S h(x)f(x)dx,$$

as soon as $\int_S h(x)f(x)dx < \infty$.

∎

Lemma 3.3 Suppose that f is bounded and positive on every compact subset of its support $\mathcal{E}_f$. If there are positive numbers ϵ and δ such that

$$q(y \mid x) > \epsilon \quad \text{for } |x - y| < \delta,$$

then the Metropolis-Hastings Markov chain is f-irreducible and aperiodic.

∎

4 Examples

4.1 Simulating the Standard Normal Distribution

This a formal demonstration of the Metropolis-Hastings algorithm. Our target distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

We are going to use a proposal random walk, where

$$Y_{n+1} = X_n + Z_{n+1},$$

and Z_n is I.I.D. $\sim U(-a, a)$. This gives us

$$q(y|x) = \begin{cases} \frac{1}{2a} & |x-y| \leq a \\ 0 & \text{otherwise.} \end{cases}$$

Cleary here the conditions of proposition 3.1 are violated, since the support of q is actually included in the support of f, not vice versa as in proposition3.1.

This is a symmetric proposal density. Hence the algorithm here is a Metropolis algorithm, where the acceptance probability only depends on the ratio $f(y)/f(x)$. Formally this is verified as follows:

$$\rho(x,y) = \begin{cases} \min\left[1, \frac{f(y)q(x|y)}{f(x)q(y|x)}\right] & |x-y| \leq a \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \min\left[1, \frac{e^{-\frac{y^2}{2}}}{e^{-\frac{x^2}{2}}}\right] & |x-y| \leq a \\ 0 & \text{otherwise} \end{cases}$$

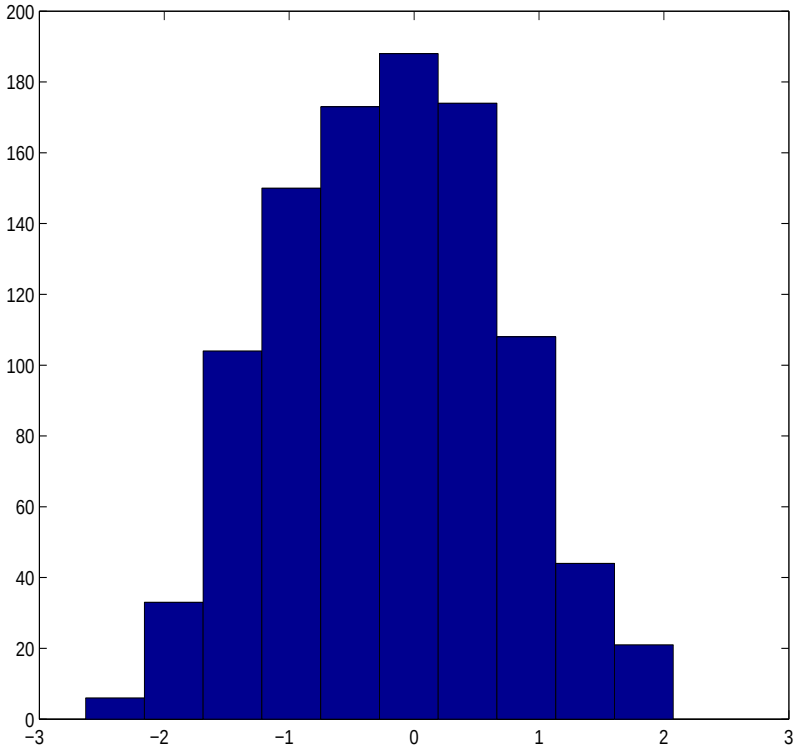
$$= \begin{cases} \min\left[1, e^{-\frac{\left(y^2-x^2\right)}{2}}\right] & |x-y| \leq a \\ 0 & \text{otherwise.} \end{cases}$$

This is impelemnted in the following Matlab^R code:

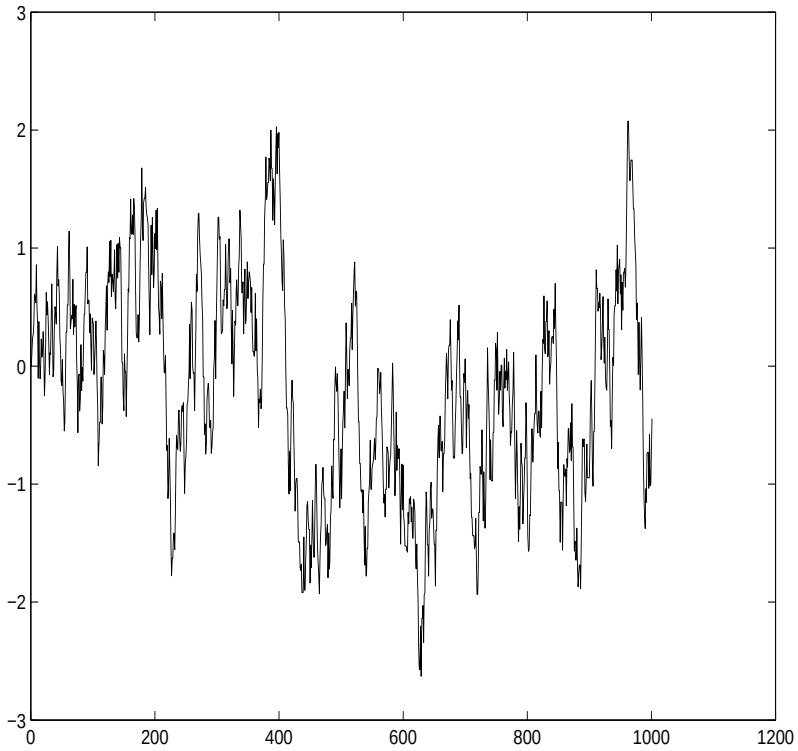
```
function x=metropolisgauss(antal,a,start);
x=start;
last=start;
for i=1:antal
y= last-a+rand*2*a;
```

```
alfa= min(1, exp(0.5 *(last^2-y^2)));
if rand(1)< alfa
x=[x y];
last=y;
else,
x=[x last];
end,
end,
```

The results are here clearly influenced by the choice of a, as examined in the Figures 1−6.



Figur 1: Histogram for the Metropolis MC with N(0,1) as the target distribution and with $a = 0.5$



Figur 2: A simulated path of the Metropolis MC with N(0,1) as the target distribution and with $a = 0.5$.

4.2 Sampling from a Posterior Density for the Parameters of a Weibull distribution

This is **example 6.3.2** in (Robert 2001, p. 305). We are dealing with the Weibull[1] distribution with the parameters

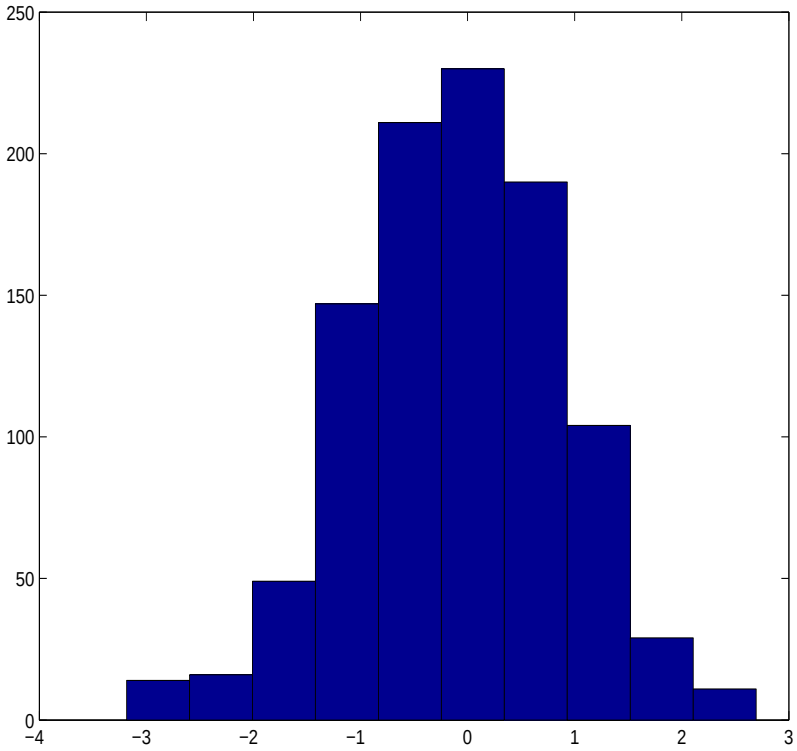
$$\theta = (\alpha, \eta)$$

in the density [2]

$$l(x|\theta) = \alpha\eta x^{\alpha-1} e^{-x^{\alpha}\eta}, x \geq 0. \tag{4.1}$$

[1] W.A. Weibull: *A Statistical Theory of the Strength of Materials*. Ingenjörsvetenskapsakademiens Handlingar No 153, Stockholm, 1939

[2] The standard representation is $l(x|\theta) = \alpha\lambda^{\alpha} x^{\alpha-1} e^{-(x\lambda)^{\alpha}}$.



Figur 3: Histogram for the Metropolis MC with N(0,1) as the target distribution and with $a = 2$

This the density of the Weibull distribution $\mathrm{We}(\alpha, \eta)$ and does not belong to the exponential family, and hence there is no explicit posterior density for θ. Let us take the prior density as

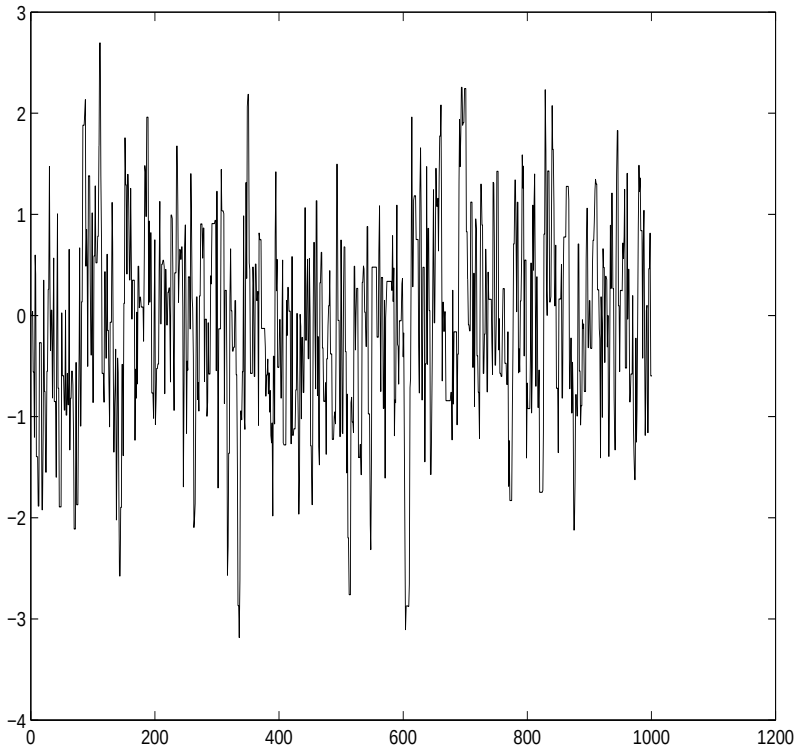
$$\phi(\theta) = C_p e^{-\alpha} \eta^{\beta-1} e^{-\xi\eta}$$

For n I.I.D. samples $x_1, \ldots, x_n \sim \mathrm{We}(\alpha, \eta)$ in (4.1) the posterior is

$$\phi(\theta|x_1, \ldots, x_n) = C_\phi \alpha^n \eta^n \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^{n} x_i^\alpha \eta} e^{-\alpha} \eta^{\beta-1} e^{-\xi\eta} \tag{4.2}$$

where C_ϕ is the numerator given by

$$C_\phi = \frac{1}{\int \prod_{i=1}^{n} l(x_i|\theta)\phi(\theta)\, d\theta},$$



Figur 4: A simulated path of the Metropolis MC with N(0,1) as the target distribution and with $a = 2$.

and cannot be given an explicit expression.

We want to explore $\phi(\theta|x_1, \ldots, x_n)$ by generating samples. Robert (loc.cit.) suggests the proposal kernel

$$q\left(\theta^{'}|\theta\right) = \frac{1}{\alpha}e^{-\frac{\alpha^{'}}{\alpha}} \cdot \frac{1}{\eta}e^{-\frac{\eta^{'}}{\eta}}. \tag{4.3}$$

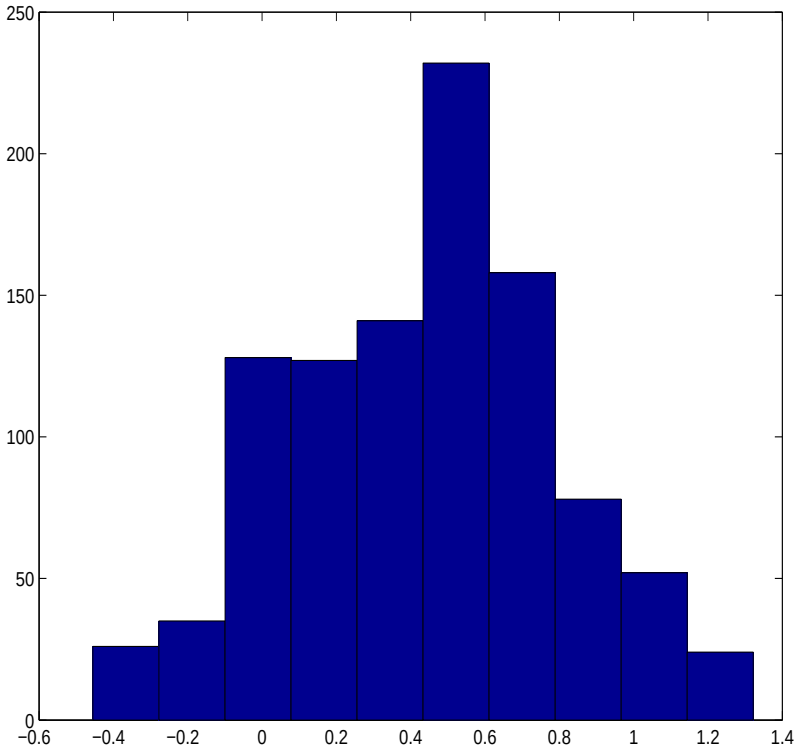
Here

$$\theta = (\alpha, \eta)$$

is current value and

$$\theta^{'} = \left(\alpha^{'}, \eta^{'}\right)$$

is the proposed value. The proposal kernel in (4.3) is a product of two expo-



Figur 5: Histogram for the Metropolis MC with N(0,1) as the target distribution and with $a = 0.1$

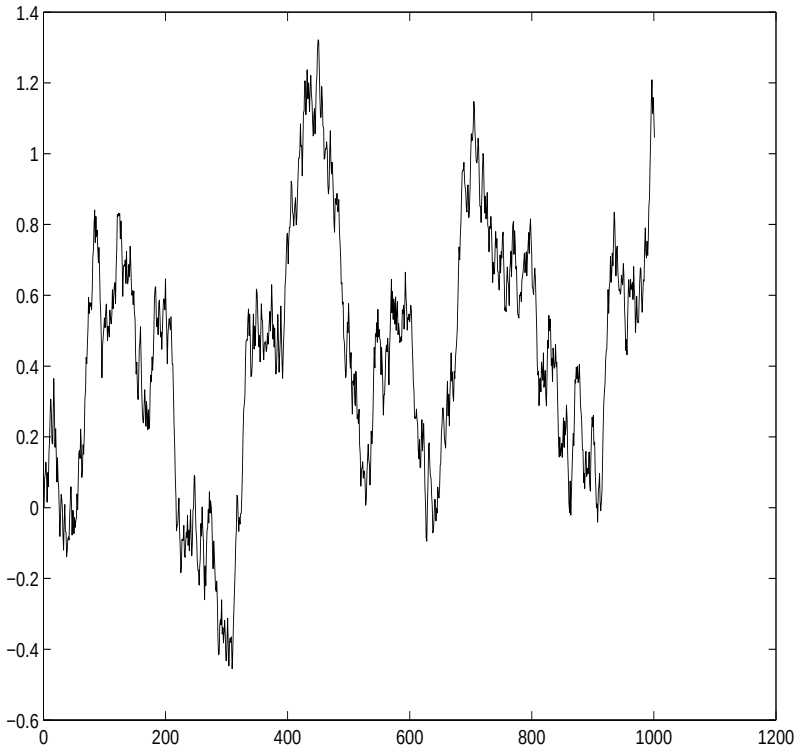
nential densities. Pseudorandom numbers drawn from an exponential distribution are obtained, e.g., by means of `exprnd` in Matlab^R Statistics Toolbox.

We can compute a as follows[3] view of (3.1).

$$\rho(\theta, \theta') = \min\left\{1, \frac{\phi\left(\theta'|x_1, \ldots, x_n\right) q\left(\theta|\theta'\right)}{\phi\left(\theta|x_1, \ldots, x_n\right) q\left(\theta'|\theta\right)}\right\}$$

$$= \min\left\{1, \frac{(\alpha')^n(\eta')^n \prod_{i=1}^n x_i^{\alpha'-1} e^{-\sum_{i=1}^n x_i^{\alpha'}\eta'} e^{-\alpha'}(\eta')^{\beta-1} e^{-\xi\eta'} \frac{1}{\alpha'\eta'} e^{-\frac{\alpha}{\alpha'}-\frac{\eta}{\eta'}}}{\alpha^n\eta^n \prod_{i=1}^n x_i^{\alpha-1} e^{-\sum_{i=1}^n x_i^{\alpha}\eta} e^{-\alpha}\eta^{\beta-1} e^{-\xi\eta} \frac{1}{\alpha\eta} e^{-\frac{\alpha'}{\alpha}-\frac{\eta'}{\eta}}}\right\}.$$

[3]It seems like there are typing errors in this on loc.cit..



Figur 6: A simulated path of the Metropolis MC with N(0,1) as the target distribution and with $a = 0.1$.

$$= \min\left\{1, \prod_{i=1}^{n} x_i^{\alpha'-\alpha}\left(\frac{\alpha'}{\alpha}\right)^{n-1}\left(\frac{\eta'}{\eta}\right)^{\beta-2+n} e^{-\xi\left(\eta'-\eta\right)} e^{\sum_{i=1}^{n} x_i^{\alpha}\eta - \sum_{i=1}^{n} x_i^{\alpha'}\eta'} e^{-(\alpha'-\alpha)} e^{-\frac{\alpha}{\alpha'}-\frac{\eta}{\eta'}+\frac{\alpha'}{\alpha}+\frac{\eta'}{\eta}}\right\}.$$

Here the inextricable constant C_ϕ is cancelled and the acceptance probability is easily implemented using some software platform like Matlab^R.

5 Implementation Issues

5.1 Various proposal densities

In order to implement the Metropolis-Hastings algorithm it is necessary to Specify a suitable proposal distribution. Typically this distribution is selected

from a family of distributions that requires the tuning of such parameters as location and scale.

One family of proposal distributions appeared already in the work of Metropolis. Then

$$q(y|x) = g_1(\|y - x\|)$$

where g_1 is a multivariate density and $\|\cdot\|$ is the Euclidean norm. The proposal is thus drawn according to the random walk

$$Y_n = X_n + Z_n,$$

where the I.I.D. sequence $Z_n \sim g_1$. Possible choices of g_1 are here the multivariate normal density and the multivariate $\mathbf{t}$ -density, where the parameters are tuned as discussed below. Another well known choice is

$$q(y|x) = g_2(y)$$

where g_2 is a multivariate density. In other words, the proposals are drawn independently of the current state, and the resulting chain is called the *independence chain*. As above, g_2 can be multivariate normal or t.

A third alternative is to use information about $f(x)$, if available (see Chib and Greenberg (1995)).

A fourth alternative is the autoregressive chain, (see Chib and Greenberg (1995)).

5.2 Tuning of the scale

The critical issue is the selection of the scale or the spread of the proposal distribution. This is an important matter that has consequences for the efficiency of the algorithm. The spread of the proposal density affects the the behaviour of the algorithm in at least two respects: one is acceptance rate (the percentage of times a move to a new point is made), and the other is the region of the sample space that is covered the chain.

To explain this, assume that the chain has converged, and the density is being sampled around its mode. Then, if the spread is very large, there are proposed candidates that will be very far from the current value, and will therefore have a low value of being accepted (because $f(y) << f(x)$). But if the spread is chosen too low, the chain will no longer traverse the support of the density, and low probability regions will be under-sampled.

6 Exercises

1. (a) n I.I.D. samples $x_1, \ldots, x_n$ of a Weibull distribution can, e.g., be obtained by using the Matlab^R Statistics Toolbox function `weibrnd`. Note the discrepancies w.r.t. the parameters in section 4.2 above and those in Matlab^R Statistics Toolbox.

   (b) Explore the posterior density in section 4.2 by Metropolis-Hastings using the prior density and proposal kernel in loc.cit. Use k independent realizations $x_1^{(l)}, \ldots, x_n^{(l)}$, $l = 1, \ldots, k$ of the Metropolis-Hastinsg chain, retain the final state $x_n^{(l)}$ and compute the histogram with these.

   (c) The function `weibfit` gives you the ML-estimate of the parameters using $x_1, \ldots, x_n$. Compare MLE with your histogram of the posterior. (Keep in mind the discrepancies in parametric representations in section 4.2 above and in Matlab^R Statistics Toolbox.)

2. Exercise 6.2.1 d. p.325 Robert (2001)

3. Exercise 6.2.6 b. & c. p.326 Robert (2001) (implement the algorithm and do simulations, the mathematics in part a. are beyond the level of attention in this course.)

7 Appendix A: A Proof of Proposition 3.1

For any $B \in \mathcal{B}(S)$ the proposal kernel is given by

$$Q(x, B) = P\left(Y_{n+1} \in B | X_n = x\right) = \int_B q(y|x)dy.$$

We define also

$$Q^{\mathbf{a}}(x, B) = P\left(Y_{n+1} \in B, Y_{n+1} \text{ is accepted } | X_n = x\right)$$

$$= \int_B q(y|x)\rho(x, y)dy. \tag{A.1}$$

which is the conditional probability that Y_{n+1} is in B and Y_{n+1} is accepted, given that $X_n = x$. We introduce also

$$s(x) = P\left(Y_{n+1} \text{ is rejected } | X_n = x\right)$$

as the conditional probability of rejecting the proposal given that $X_n = x$.

We shall next derive the transition kernel $P\left(X_{n+1} \in B | X_n = x\right)$. By the law of total probability

$$P\left(X_{n+1} \in B | X_n = x\right) =$$

$$= P\left(X_{n+1} \in B, Y_{n+1} \text{ is accepted } | X_n = x\right) +$$

$$P\left(X_{n+1} \in B, Y_{n+1} \text{ is rejected and } x \in B \ | X_n = x\right)$$

$$= Q^{\mathbf{a}}(x, B) + I_B(x)s(x), \tag{A.2}$$

where $I_B(x)$ is the indicator function of B.

Now we show that the density $f(x)$ is the density of the invariant distribution with respect the kernel in (A.2) above. We show that the detailed balance condition holds via checking (2.13). We have

$$P\left(X_n \in A, X_{n+1} \in B\right) = \int_A f(x) P\left(X_{n+1} \in B | X_n = x\right) dx$$

and from (A.2) and (A.1) we get

$$= \int_A f(x) Q^{\mathbf{a}}(x, B)dx + \int_A f(x) I_B(x) s(x)dx$$

$$= \int_A f(x) \left(\int_B q(y|x)\rho(x,y)dy \right) dx + \int_{A\cap B} f(x)s(x)dx \qquad \text{(A.3)}$$

Let us now define for any $x \in A$

$$B(x) = \{y \in B | f(y)q(x|y) > f(x)q(y|x)\}.$$

We have

$$B = B(x) \bigcup B(x)^c.$$

By (3.1) we have in $B(x)$ that

$$f(y)q(x|y) > f(x)q(y|x) \Leftrightarrow \rho(x,y) = \min\left\{1, \frac{f(y)q(x|y)}{f(x)q(y|x)}\right\} = 1. \qquad \text{(A.4)}$$

We have

$$\begin{aligned} \int_A f(x) \left(\int_B q(y|x)\rho(x,y)dy \right) dx &= \int_A f(x) \left(\int_{B(x)} q(y|x)\rho(x,y)dy \right) dx \\ &\quad + \int_A f(x) \left(\int_{B(x)^c} q(y|x)\rho(x,y)dy \right) dx. \end{aligned} \qquad \text{(A.5)}$$

In the first integral on the right hand side

$$\int_A f(x) \left(\int_{B(x)} q(y|x)\rho(x,y)dy \right) dx = \int_A f(x) \left(\int_{B(x)} q(y|x)dy \right) dx.$$

But under (A.4) we have also

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

so that we get

$$\int_A f(x) \int_{B(x)} q(y|x)dydx = \int_A \int_{B(x)} f(y)q(x|y)\rho(y,x)dydx. \qquad \text{(A.6)}$$

In the second integral in (A.5) we get

$$\int_A f(x) \left(\int_{B(x)^c} q(y|x)\rho(x,y)dy \right) dx$$

$$= \int_A f(x) \left(\int_{B(x)^c} q(y|x) \frac{f(y)q(x|y)}{f(x)q(y|x)} dy \right) dx$$

$$= \int_A \left(\int_{B(x)^c} f(y)q(x|y)dy \right) dx.$$

$$= \int_A \left(\int_{B(x)^c} f(y)q(x|y)\rho(y,x)dy \right) dx.$$

since in $B(x)^c$ we have that

$$\rho(y,x) = \min\left\{1, \frac{f(x)q(y|x)}{f(y)q(x|y)}\right\} = 1.$$

Therefore we get in view of (A.5) that

$$\int_A f(x) \left(\int_B q(y|x)\rho(x,y)dy \right) dx + \int_{A\cap B} f(x)s(x)dx \tag{A.7}$$

$$= \int_A \left(\int_B f(y)q(x|y)\rho(y,x)dy \right) dx + \int_{A\cap B} f(x)s(x)dx. \tag{A.8}$$

Then we have

$$\int_A \left(\int_B f(y)q(x|y)\rho(y,x)dy \right) dx = \int_B f(y) \int_A q(x|y)\rho(y,x)dxdy.$$

We have

$$\int_B f(y) \int_A q(x|y)\rho(y,x)dxdy = \int_B f(y)Q^{\mathbf{a}}(y,A)dx. \tag{A.9}$$

The second term in (A.3) is

$$\int_{A\cap B} f(x)s(x)dx = \int_B I_A(x)f(x)s(x)dx = \int_B I_A(y)f(y)s(y)dy$$

by the change of variable $x \mapsto y$, the Jacobian of which is 1. By insertion of this and (A.9) in (A.3) we get

$$P\left(X_n \in A, X_{n+1} \in B\right) = \int_B f(y)Q^{\mathbf{a}}(y,A)dx + \int_B I_A(y)f(y)s(y)dy$$

$$= P\left(X_n \in B, X_{n+1} \in A\right),$$

which establishes, in view of (A.2), the reversibility condition (2.13) in the case (A.4). Therefore, due to proposition 2.7, f is the invariant density of the Metropolis-Hastings chain.

The assumption about the inclusion of pertinent supports are brodaly speaking needed in order to guarantee that, e.g., an integral like

$$\int_A f(x) \left(\int_B q(y|x)\rho(x,y)dy \right) dx$$

does not vanish over a set B, where the target density is positive. ∎

8 References and further reading:

1 Journal articles and technical reports:

- S. Chib and E. Greenberg (1995): Understanding the Metropolis-Hastings Algorithm. *The American Statistician* , 49, pp. 327−355.
- W.K. Hastings (1970): Monte Carlo sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, pp. 97−109.
- L. Tierney (1994): Markov chains for exploring posterior distributions. *Annals of Statistics*, 22, pp. 1701−1762.

2 Books:

- G. Englund (2000): *Datorintensiva metoder i matematisk statistik*. Institutionen för matematik, KTH, Stockholm.
- S.P. Meyn and R. Tweedie (993): *Markov Chains and Stochastic Stability*. Springer Verlag. London, Berlin, e.t.c.
- E. Nummelin (1984): *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press.
- C.P. Robert (2001): *The Bayesian Choice. Second Edition*. Springer Verlag, New York.
- C.P. Robert and G. Casella (1999): *Monte Carlo Statistical Methods*. Springer Verlag, New York.