Computational statistics

Markov Chain Monte Carlo methods

Thierry Denœux

March 2017

Overview

Contents of this chapter

- When a target density f can be evaluated but not easily sampled, the methods from the previous chapter can be applied to obtain an approximate or exact sample. The primary use of such a sample is to estimate the expectation of a function of $X \sim f(x)$.
- The Markov chain Monte Carlo (MCMC) methods introduced in this chapter can also be used to generate a draw from a distribution that approximates f and estimate expectations of functions of X.
- MCMC methods are distinguished from the simulation techniques in the previous chapter by their iterative nature and the ease with which they can be customized to very diverse and difficult problems.

Basic ideas

- Let the sequence $\{X^{(t)}\}$ denote a Markov chain for $t = 0, 1, 2, \ldots$, where $X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ and the state space is either continuous or discrete.
- The MCMC sampling strategy is to construct a Markov chain that converges to a stationary distribution equal to the target distribution f.
- For sufficiently large t, a realization $X^{(t)}$ from this chain will have approximate marginal distribution f .
- A very popular application of MCMC methods is to facilitate Bayesian inference where f is a Bayesian posterior distribution for parameters X.
- The art of MCMC lies in the construction of a suitable chain.

Overview

Notations

- Consider a sequence of random variables $\{X^{(t)}\}$, $t = 0, 1, \ldots$, where each $X^{(t)}$ may equal one of a finite or countably infinite number of possible values, called states.
- The notation $X^{(t)} = j$ indicates that the process is in state j at time t.
- The state space, $\mathcal{S}$, is the set of possible values of the random variable $X^{(t)}$.

Markov property

- The joint distribution of $X^{(0)}, \ldots, X^{(t)}$ can be written as the product of conditional distributions of each random variable given its history,

$$p(x^{(0)}, \ldots, x^{(n)}) = p(x^{(n)} \mid x^{(0)}, \ldots, x^{(n-1)}) \times p(x^{(n-1)} \mid x^{(0)}, \ldots, x^{(n-2)}) \times \ldots \times p(x^{(1)} \mid x^{(0)}) p(x^{(0)}). \quad (1)$$

- The sequence $\{X^{(t)}\}$, $t = 0, 1, \ldots$, is a Markov chain (MC) if

$$p(x^{(t)} \mid x^{(0)}, \ldots, x^{(t-1)}) = p(x^{(t)} \mid x^{(t-1)})$$

  for all t and all $x^{(0)}, \ldots, x^{(t)}$.
- Then, (1) can be simplified to

$$p(x^{(0)}, \ldots, x^{(n)}) = p(x^{(0)}) \prod_{t=1}^{n} p(x^{(t)} \mid x^{(t-1)}). \quad (2)$$

Transition probabilities

- Let $p_{ij}^{(t)}$ be the probability that the observed state changes from state i at time t to state j at time $t+1$,

$$p_{ij}^{(t)} = P(X^{(t+1)} = j \mid X^{(t)} = i)$$

- The quantity $p_{ij}^{(t)}$ is called the one-step transition probability.
- If none of the one-step transition probabilities change with t, then the MC is called time-homogeneous, and $p_{ij}^{(t)} = p_{ij}$. If any of the one-step transition probabilities change with t, then the MC is called time-inhomogeneous.

Transition probability matrix

- A time-homogeneous MC is governed by a transition probability matrix.
- Suppose there are s states in $\mathcal{S}$. Then matrix $\boldsymbol{P} = (p_{ij})$ of size $s \times s$ is called the transition probability matrix.
- Each element in P must be between zero and one, and each row of the matrix must sum to one, as

$$\sum_{j=1}^{s} p_{ij} = 1.$$

We say that $\boldsymbol{P}$ is a stochastic matrix.

Definitions

- A MC is irreducible if any state j can be reached from any state i in a finite number of steps for all i and j. In other words, for each i, j and n there must exist $m > 0$ such that

$$P[X^{(m+n)} = j \mid X^{(n)} = i] > 0.$$

- A MC is periodic if it can visit certain portions of the state space only at certain regularly spaced intervals. State j has period d if the probability of going from state j to state j in n steps is 0 for all n not divisible by d.
- If every state in a MC has period 1, then the chain is called aperiodic.

Stationary distribution

- Let π denote a vector of probabilities that sum to one, with i-th element π_i denoting the marginal probability that $X^{(t)} = i$.
- Then the marginal distribution of $X^{(t+1)}$ is

$$P[X^{(t+1)} = j] = \sum_{i=1}^{s} P(X^{(t+1)} = j \mid X^{(t)} = i) P[X^{(t)} = i]$$
$$= \sum_{i=1}^{s} p_{ij} \pi_i = [\pi^T \boldsymbol{P}]_j.$$

- Any discrete probability distribution π such that $\pi^T \boldsymbol{P} = \pi^T$ is called a stationary distribution for $\boldsymbol{P}$, or for the MC having transition probability matrix $\boldsymbol{P}$.
- If $\{X^{(t)}\}$ follows a stationary distribution, then the marginal distributions of $X^{(t)}$ and $X^{(t+1)}$ are identical.

Example

- Let

$$\boldsymbol{P} = \begin{pmatrix} 0.75 & 0.25 \\ 0.125 & 0.875 \end{pmatrix}$$

- Does $\boldsymbol{P}$ have a stationary distribution?
- Let $\pi = (\pi_1, 1 - \pi_1)$. It is stationary iff $\pi^T \boldsymbol{P} = \pi^T$. We get the equation

$$0.75\pi_1 + 0.125(1 - \pi_1) = \pi_1 \Leftrightarrow = \pi_1 = 1/3.$$

- The unique solution is $\pi = (1/3, 2/3)^T$.

Theorem

- If a MC with transition probability matrix $\boldsymbol{P}$ and stationary distribution π is irreducible and aperiodic, then π is unique and for all i,

$$\lim_{t \to \infty} P[X^{(t+1)} = j \mid X^{(t)} = i] = \pi_j.$$

- The π_j are the solutions of the following set of equations:

$$\pi_j \geq 0, \quad \sum_i \pi_i = 1, \quad \pi_j = \sum_i \pi_i p_{ij}, \ \forall j$$

Ergodic theorem

- If $\{X^{(t)}\}$ is an irreducible and aperiodic MC with stationary distribution π, then $X^{(t)}$ converges in distribution to the distribution given by π, and for any function h,

$$\frac{1}{n}\sum_{t=1}^{n} h(X^{(t)}) \to \mathbb{E}_\pi\{h(X)\}$$

almost surely as $n \to \infty$, provided $\mathbb{E}_\pi\{h(X)\}$ exists.

Continuous state spaces

- Similar results hold for continuous state spaces.
- In the continuous case, a time-homogeneous MC is defined by the transition kernel

$$f(x, x') = f_{X^{(t+1)}|X^{(t)}}(x' \mid x),$$

so that

$$f(x^{(0)}, \ldots, x^{(n)}) = f(x^{(0)}) \prod_{t=1}^{n} f(x^{(t-1)}, x^{(t)})$$

- The density π is stationary for the MC with kernel $f(x, x')$ is

$$\pi(x') = \int f(x, x')\pi(x)dx.$$

Asymptotic results

- Under similar conditions as in the finite case, we have, for a stationary density π,

$$(X^{(t)}) \xrightarrow{d} \pi$$

and

$$\frac{1}{n}\sum_{t=1}^{n} h(X^{(t)}) \xrightarrow{a.s.} \mathbb{E}_{\pi}\{h(X)\} \tag{3}$$

Overview

Metropolis-Hastings (MH) algorithm

- A very general method for constructing a MC.
- The method begins at $t = 0$ with the selection of $X^{(0)} = x^{(0)}$ drawn from some starting distribution g, with the requirement that $f(x^{(0)}) > 0$. Given $X^{(t)} = x^{(t)}$, we generate $X^{(t+1)}$ as follows:
  1. Sample a candidate value X^* from a proposal distribution $g(\cdot \mid x^{(t)})$.
  2. Compute the MH ratio $R(x^{(t)}, X^*)$ with

$$R(u, v) = \frac{f(v)g(u \mid v)}{f(u)g(v \mid u)}$$

  3. Sample a value for $X^{(t+1)}$ according to the following:

$$X^{(t+1)} = \begin{cases} X^* & \text{with probability } \min[R(x^{(t)}, X^*), 1], \\ x^{(t)} & \text{otherwise.} \end{cases}$$

  4. Increment t and return to step 1.

Properties

- Clearly, a chain constructed via the MH algorithm is Markov since $X^{(t+1)}$ is only dependent on $X^{(t)}$.
- Whether the chain is irreducible and aperiodic depends on the choice of proposal distribution; the user must check these conditions for any implementation.
- If this check confirms irreducibility and aperiodicity, then the chain generated by the MH algorithm has a unique limiting stationary distribution, which is the target distribution f.

Proof

- Suppose $X^{(t)} \sim f(x)$, and consider two points in the state space of the chain, say x_1 and x_2, for which $f(x_1) > 0$ and $f(x_2) > 0$. Without loss of generality, label these points in the manner such that $f(x_2)g(x_1 \mid x_2) \geq f(x_1)g(x_2 \mid x_1)$.
- The joint density of $(X^{(t)}, X^{(t+1)})$ at (x_1, x_2) is $f(x_1)g(x_2 \mid x_1)$ because if $X^{(t)} = x_1$ and $X^* = x_2$, then $R(x_1, x_2) \geq 1$ so $X^{(t+1)} = x_2$.
- The joint density of $(X^{(t)}, X^{(t+1)})$ at (x_2, x_1) is

$$f(x_2)g(x_1 \mid x_2)\frac{f(x_1)g(x_2 \mid x_1)}{f(x_2)g(x_1 \mid x_2)} = f(x_1)g(x_2 \mid x_1)$$

because we need to start with $X^{(t)} = x_2$, to propose $X^* = x_1$, and then to set $X^{(t+1)}$ equal to X^* with probability $R(x_1, x_2)$.

Proof (continued)

- Consequently, the joint density of $(X^{(t)}, X^{(t+1)})$ is symmetric:

$$f_{(X^{(t)},X^{(t+1)})}(x_1, x_2) = f_{(X^{(t)},X^{(t+1)})}(x_2, x_1).$$

- Hence $X^{(t)}$ and $X^{(t+1)}$ have the same marginal distributions.
- Thus the marginal distribution of $X^{(t+1)}$ is f, and f must be the stationary distribution of the chain.

Application

- Recall from Equation (3) that we can approximate the expectation of a function of a random variable by averaging realizations from the stationary distribution of a MH chain.
- The distribution of realizations from the MH chain approximates the stationary distribution of the chain as t progresses; therefore $\mathbb{E}\{h(X)\} \approx \sum_{t=1}^{n} h(x^{(t)})$.
- Some of the useful quantities that can be estimated this way include means $\mathbb{E}\{h(X)\}$, variances $\mathbb{E}[h(X) - \mathbb{E}\{h(X)\}]^2$, and tail probabilities $\mathbb{E}\{I(h(X) \leq q)\}$ for some constant q.

Importance of the proposal distribution

- A well-chosen proposal distribution produces candidate values that cover the support of the stationary distribution in a reasonable number of iterations and produces candidate values that are not accepted or rejected too frequently:
  - If the proposal distribution is too diffuse relative to the target distribution, the candidate values will be rejected frequently and thus the chain will require many iterations to adequately explore the space of the target distribution.
  - If the proposal distribution is too focused (e.g., has too small a variance), then the chain will remain in one small region of the target distribution for many iterations while other regions of the target distribution will not be adequately explored.
- Next we introduce several MH variants obtained by using different classes of proposal distributions.

Overview

Independence Chains

- Suppose that the proposal distribution for the MH algorithm is chosen such that $g(x^* \mid x^{(t)}) = g(x^*)$ for some fixed density g.
- This yields an independence chain, where each candidate value is drawn independently of the past. In this case, the MH ratio is

$$R(x^{(t)}, X^*) = \frac{f(X^*)g(x^{(t)})}{f(x^{(t)})g(X^*)}.$$

- The resulting Markov chain is irreducible and aperiodic if $g(x) > 0$ whenever $f(x) > 0$.
- The proposal distribution g should resemble the target distribution f, but should cover f in the tails.

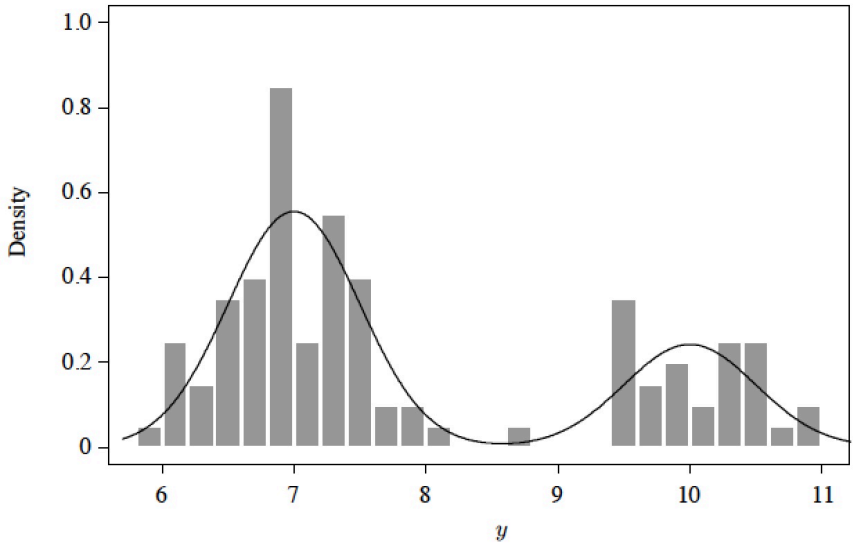
Bayesian Inference

- For Bayesian inference, a very simple strategy is to use the prior as a proposal distribution in an independence chain.
- In our MH notation, $f(\theta) = p(\theta \mid y)$ and $g(\theta^*) = p(\theta^*)$. Conveniently, this means

$$R(\theta^{(t)}, \theta^*) = \frac{p(\theta^* \mid y)p(\theta^{(t)})}{p(\theta^{(t)} \mid y)p(\theta^*)} = \frac{L(\theta^* \mid y)}{L(\theta^{(t)} \mid y)}.$$

- In other words, we propose from the prior, and the MH ratio equals the likelihood ratio.
- By construction, the support of the prior covers the support of the target posterior, so the stationary distribution of this chain is the desired posterior.

Example: Mixture Distribution



- Suppose we have observed data $y_1, y_2, \ldots, y_{100}$ iid from the mixture distribution

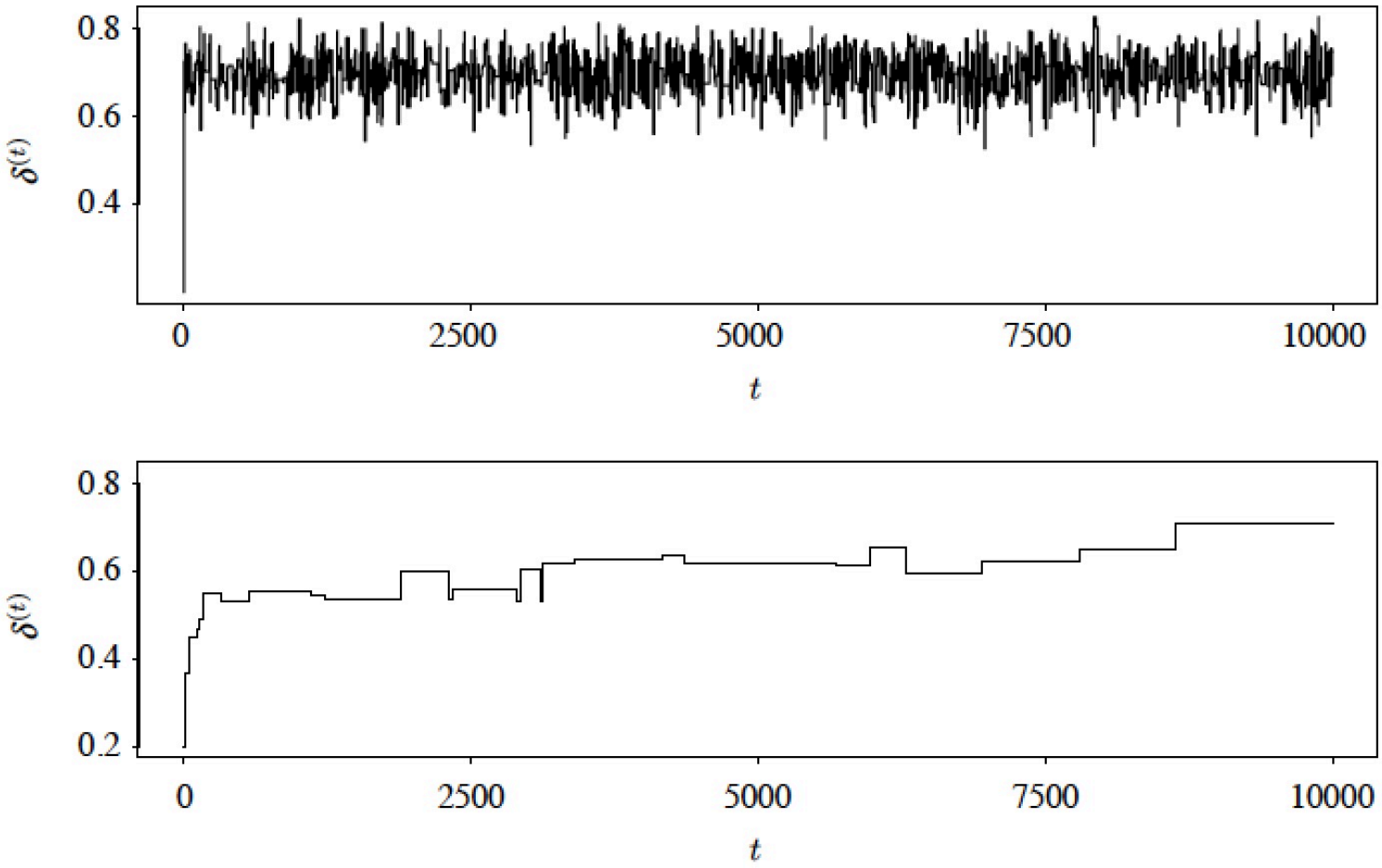
$$\delta N(7, 0.5^2) + (1 - \delta) N(10, 0.5^2)$$

- We will use MCMC techniques to construct a chain whose stationary distribution equals the posterior density of δ. The data were generated with $\delta = 0.7$, so we should find that the posterior density is concentrated in this area.

Proposal distributions

- In this example,we try two different independence chains. In the first case we use a $Beta(1,1)$ density as the proposal density, and in the second case we use a $Beta(2,10)$ density.
- The first proposal distribution is equivalent to a $Unif(0,1)$ distribution, while the second is skewed right with mean approximately equal to 0.167. In this second case values of δ near 0.7 are unlikely to be generated from the proposal distribution.
- In the next figure shows the sample paths for 10,000 iterations of both chains. A sample path is a plot of the chain realizations $\delta^{(t)}$ against the iteration number t. This plot is useful for investigating the behavior of the Markov chain and is discussed further in the sequel.

Sample paths

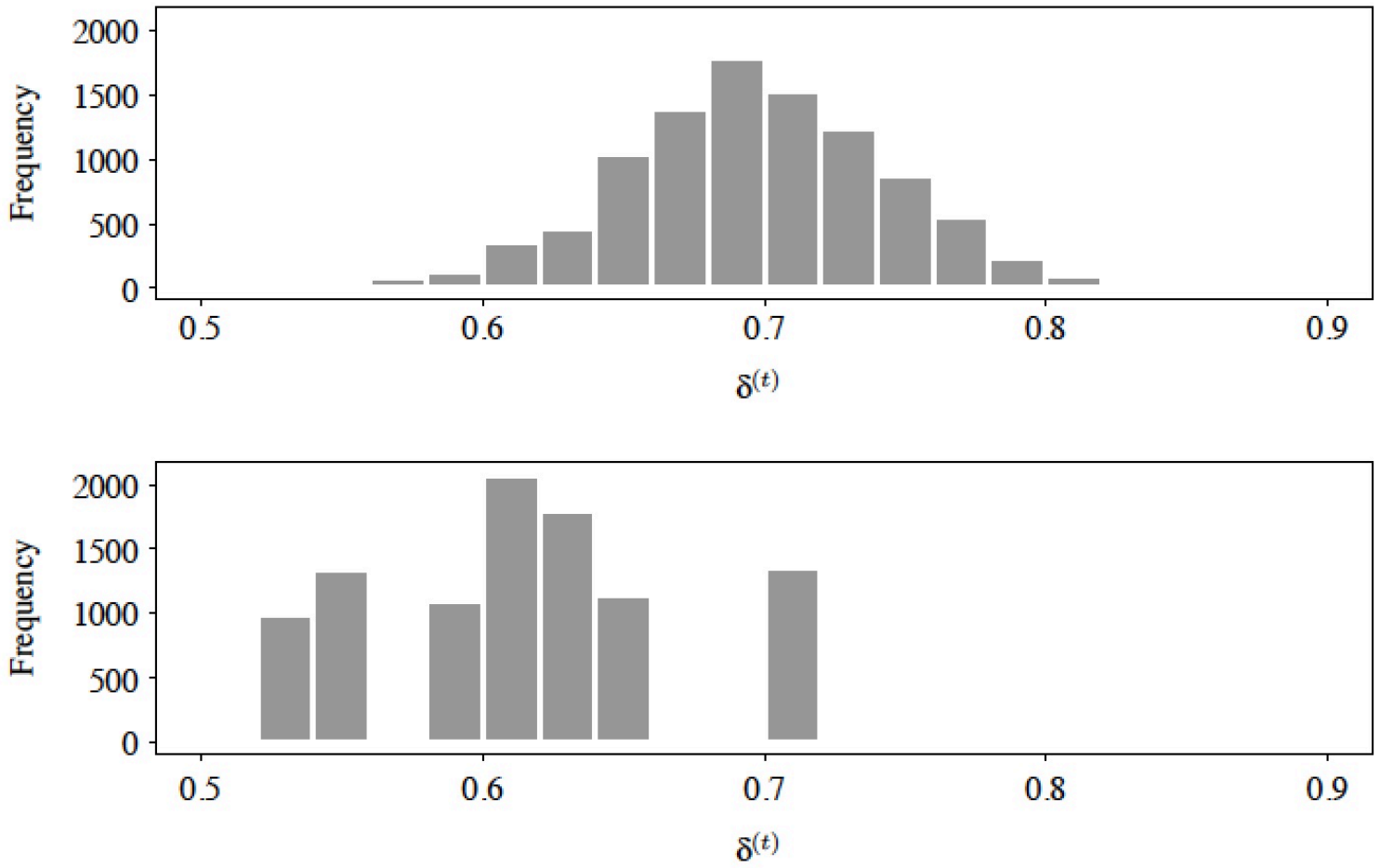


Sample paths for δ from independence chains with proposal densities $Beta(1, 1)$ (top) and $Beta(2, 10)$ (bottom).

Interpretation

- The upper panel shows a Markov chain that moves quickly away from its starting value and seems easily able to sample values from all portions of the parameter space supported by the posterior for δ. Such behavior is called good mixing.
- The lower panel corresponds to the chain using a $Beta(2, 10)$ proposal density. The resulting chain moves slowly from its starting value and does a poor job of exploring the region of posterior support (i.e., poor mixing). This chain has clearly not converged to its stationary distribution since drift is still apparent. Such a plot should make the MCMC user reconsider the proposal density.

Estimated posterior distributions



Histograms of $\delta^{(t)}$ for iterations 201-10,000 of independence chains with proposal densities $Beta(1,1)$ (top) and $Beta(2,10)$ (bottom).

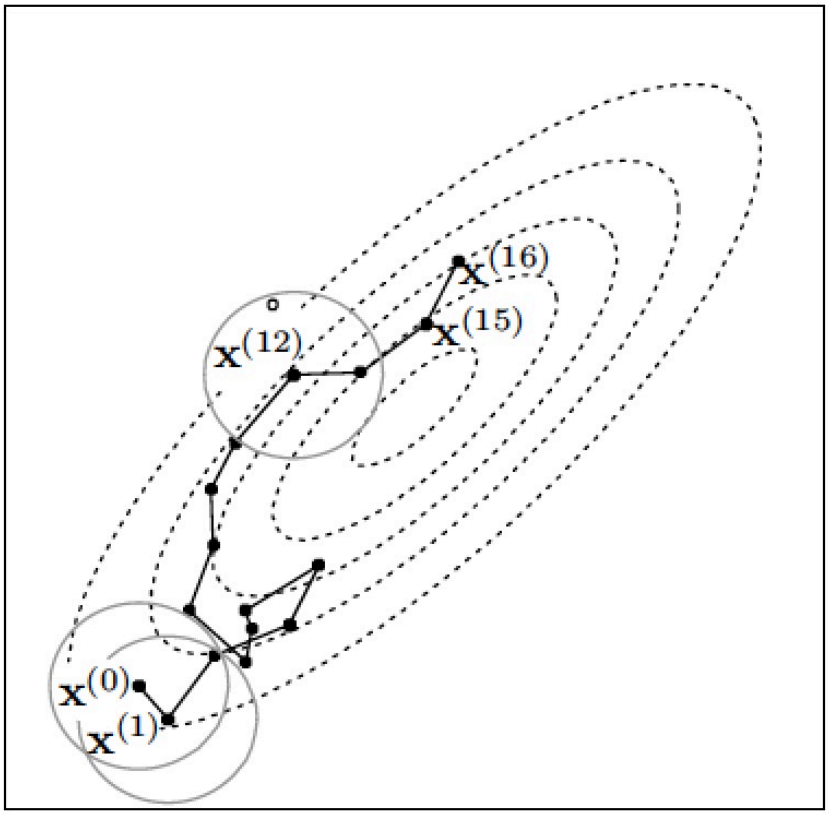
Overview

Random Walk Chains

- A random walk chain is another type of Markov chain produced via a simple variant of the MH algorithm.
- Let X^* be generated by drawing $\epsilon \sim h(\epsilon)$ for some density h and then setting $X^* = x^{(t)} + \epsilon$. This yields a random walk chain. In this case, $g(x^* \mid x^{(t)}) = h(x^* - x^{(t)})$.
- Common choices for h include a uniform distribution over a ball centered at the origin, a scaled standard normal distribution or a scaled Student's t distribution.
- If the support region of f is connected and h is positive in a neighborhood of 0, the resulting chain is irreducible and aperiodic.
- If $h(-\epsilon) = -h(\epsilon)$, the MH ratio becomes simply

$$R(x^{(t)}, X^*) = \frac{f(X^*)}{f(x^{(t)})}.$$

Random Walk Chain Example



Hypothetical random walk chain for sampling a two-dimensional target distribution (dotted contours) using proposed increments sampled uniformly from a disk centered at the current value.

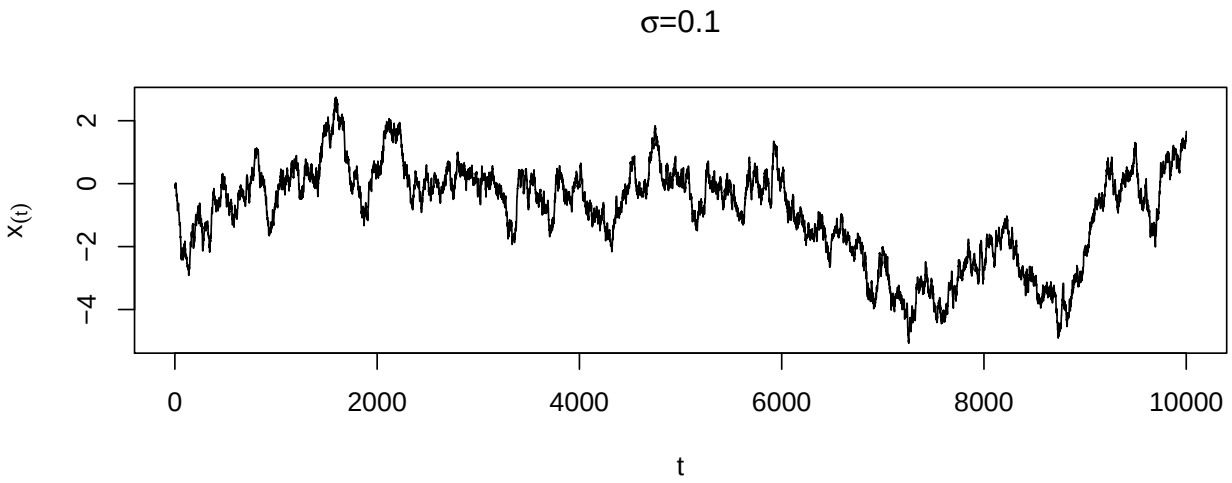
Example

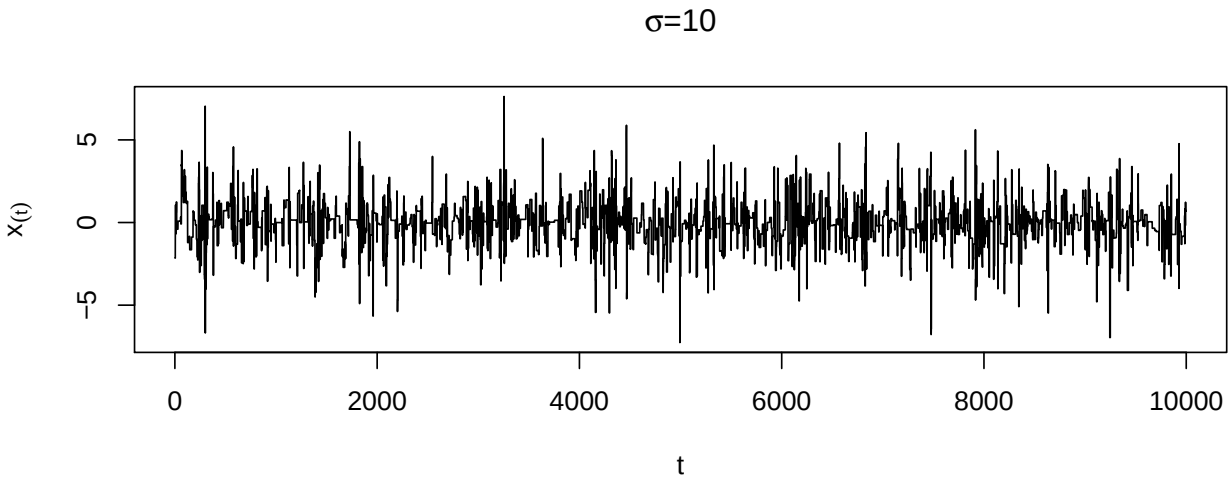
- Assume we want to construct a random walk MH sampler to generate a sample of 10,000 observations from the Laplace distribution,

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < +\infty.$$

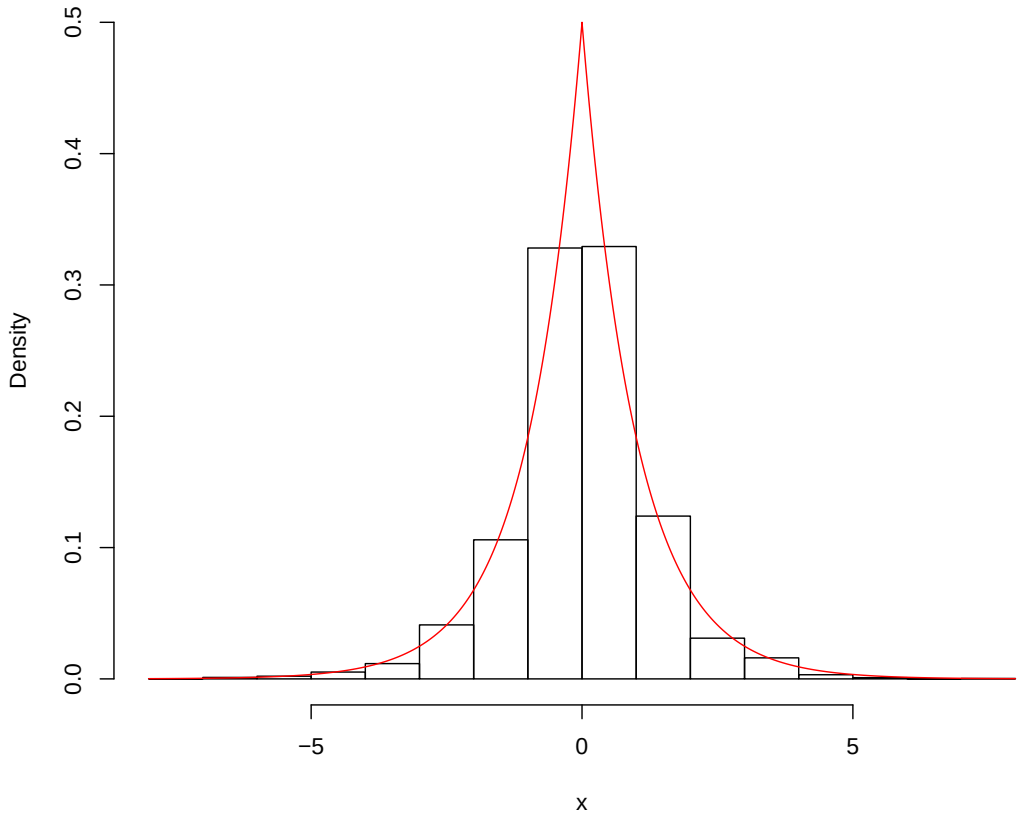
- We use a random-walk chain with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to generate proposals $X^* = x^{(t)} + \epsilon$.

Results


σ=0.1
x(t)
2
0
−2
−4
0
2000
4000
6000
8000
10000
t


σ=10
x(t)
5
0
−5
0
2000
4000
6000
8000
10000
t

Results (continued)



Histogram of simulated values from $t = 200$ to $t = 10,000$, obtained with $\sigma = 10$.

Overview

Overview