Computational statistics

Chapter 5: Markov Chain Monte Carlo methods

Thierry Denœux
Université de technologie de Compiègne

January-March 2024

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 chain

Definition

Definition

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 $x^{(0)}, \ldots, x^{(t)}$.

Markov chain

Property

For a MC, the joint distribution of $X^{(0)}, \ldots, X^{(n)}$ for any n can be expressed in a simple way:

- In general, we have

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

- For a MC, (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 $\mathbf{P} = (p_{ij})$ of size $s \times s$ is called the transition probability matrix.
- Each element in $\mathbf{P}$ 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 $\mathbf{P}$ 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. In other words, for all 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

- Consider a time-homogeneous MC. 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

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

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

Example

- Let
$$\mathbf{P} = \begin{pmatrix} 0.75 & 0.25 \\ 0.125 & 0.875 \end{pmatrix}$$
- Does $\mathbf{P}$ have a stationary distribution?
- Let $\pi = (\pi_1, 1-\pi_1)^T$. It is stationary iff $\pi^T \mathbf{P} = \pi^T$. We get the equation
$$0.75\pi_1 + 0.125(1-\pi_1) = \pi_1 \Leftrightarrow \pi_1 = 1/3.$$
- The unique solution is $\pi = (1/3, 2/3)^T$.

Important result

Theorem

If $\{X^{(t)}\}$ is an irreducible and aperiodic MC with stationary distribution π, then

1. *$X^{(t)}$ converges in distribution to a r.v. X with the distribution given by π, and*
2. *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

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

so that

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

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

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

Asymptotic result

Theorem

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

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

where X *is a r.v. with density* π, *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 $U \sim \text{Unif}(0, 1)$ and define $X^{(t+1)}$ as follows:

$$X^{(t+1)} = \begin{cases} X^* & \text{if } U \leq R(x^{(t)}, X^*) \\ x^{(t)} & \text{otherwise.} \end{cases}$$

  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 $f(x_2)g(x_1 \mid x_2) \geq f(x_1)g(x_2 \mid x_1)$, i.e., $R(x_1, x_2) \geq 1$.
- 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 $R(X^{(t)}, X^*) \geq 1$ 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 we set $X^{(t+1)}$ equal to X^* with probability $R(x_2, x_1)$.

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

$$\mathbb{E}\{h(X)\} \approx \frac{1}{n}\sum_{t=1}^{n} h(x^{(t)}).$$

- 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.

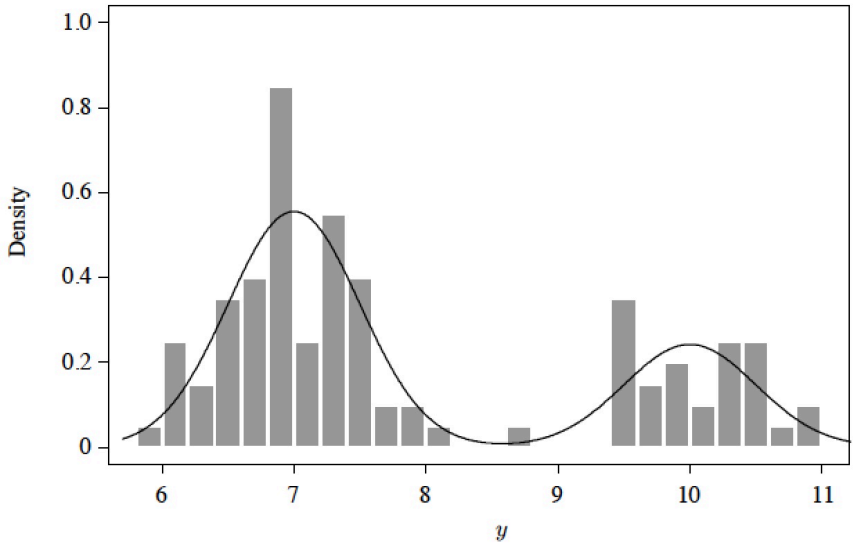
Bayesian Inference

- 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

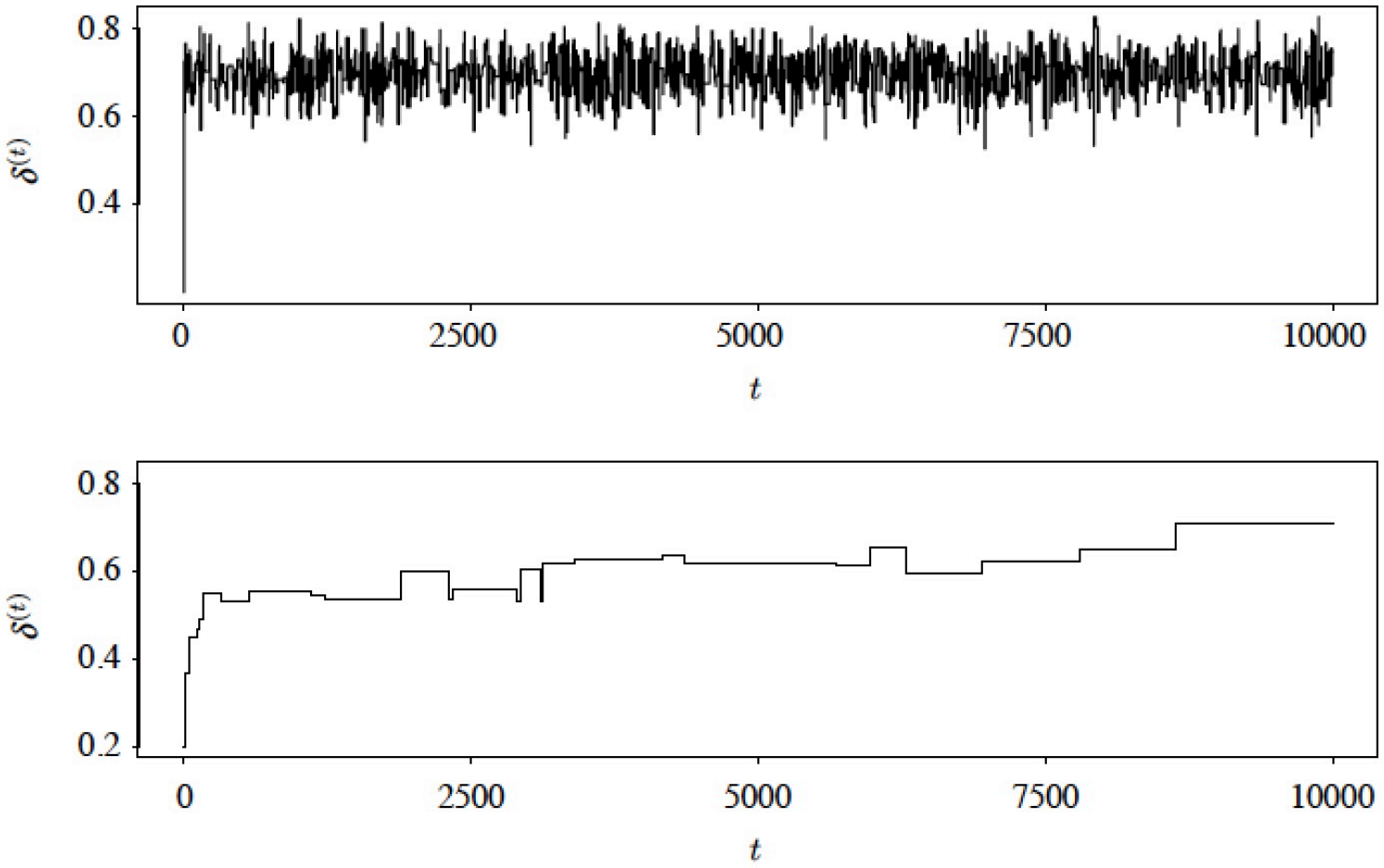
$$\delta N(7, 0.5^2) + (1 - \delta) N(10, 0.5^2)$$

- 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.
- 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

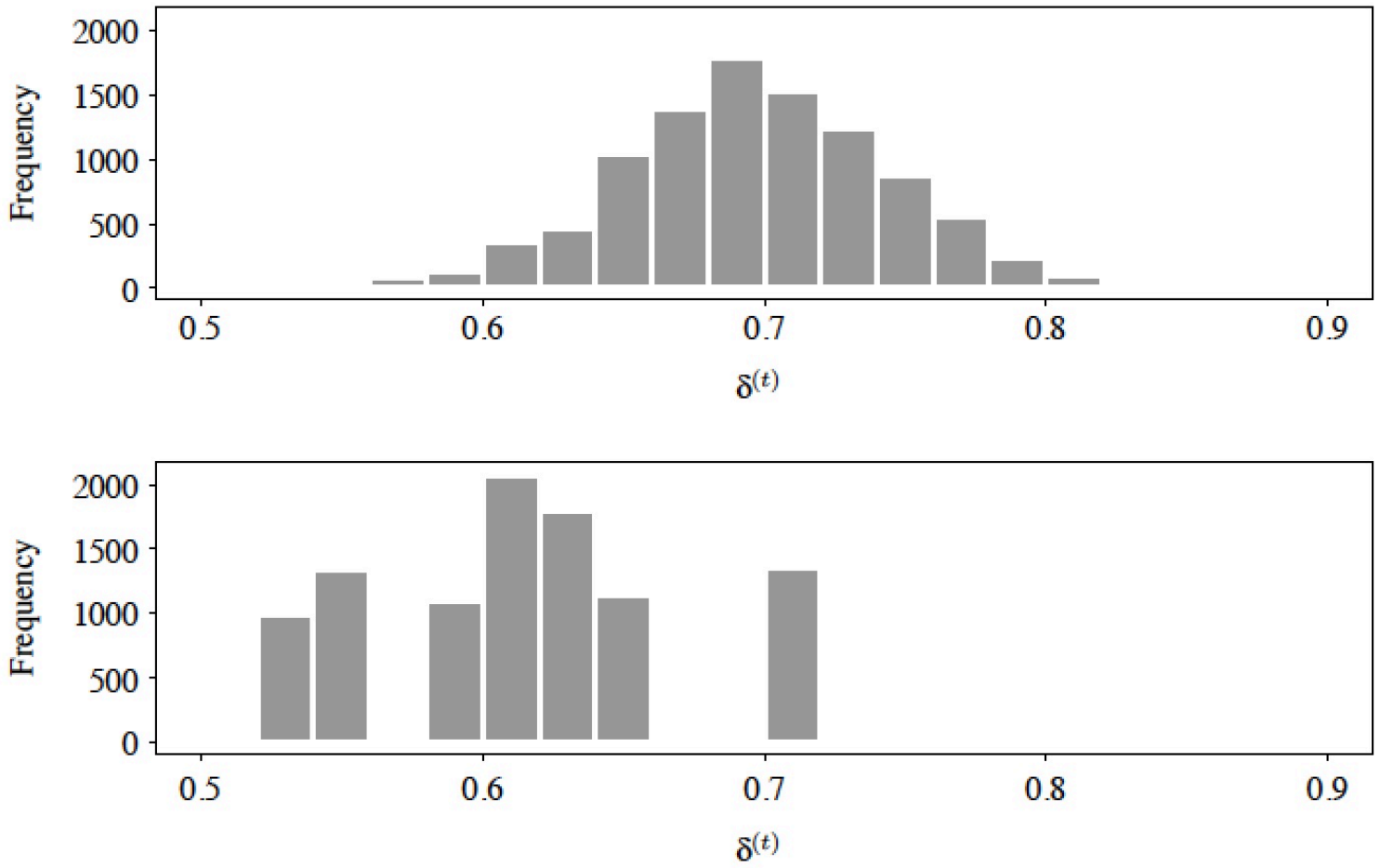


Sample paths for δ from independence chains with proposal densities $Beta(1,1)$ (top) and $Beta(2,10)$ (bottom).

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).

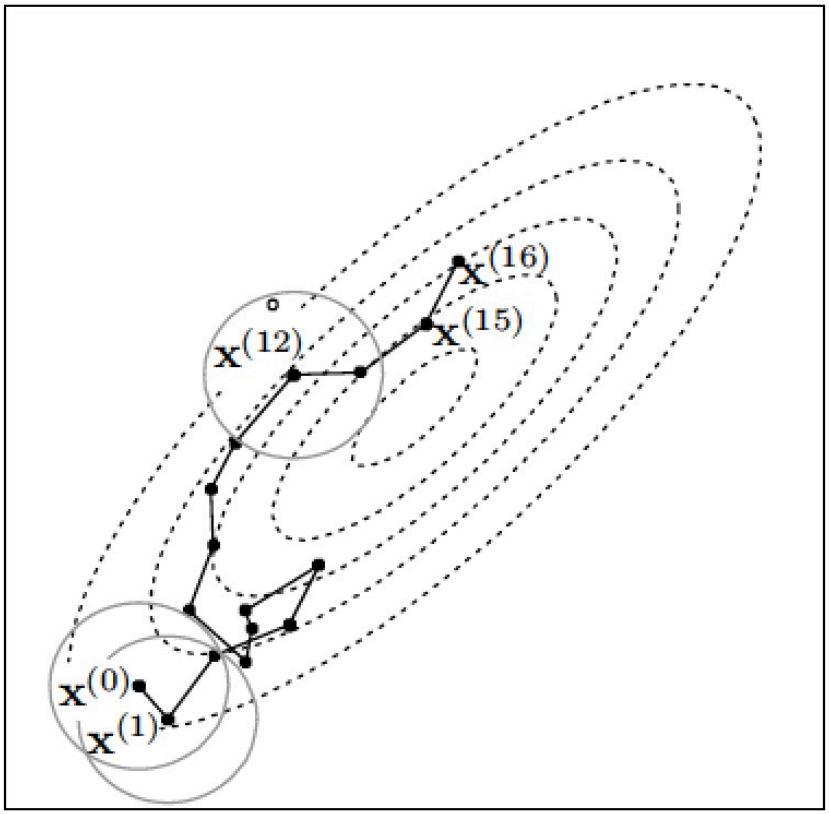
Overview

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.

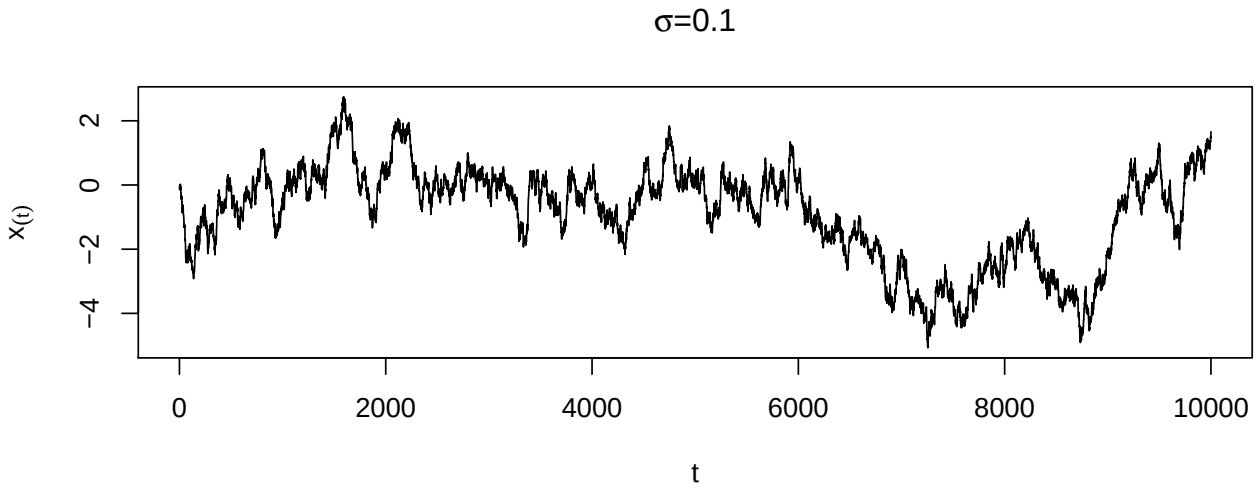
Example

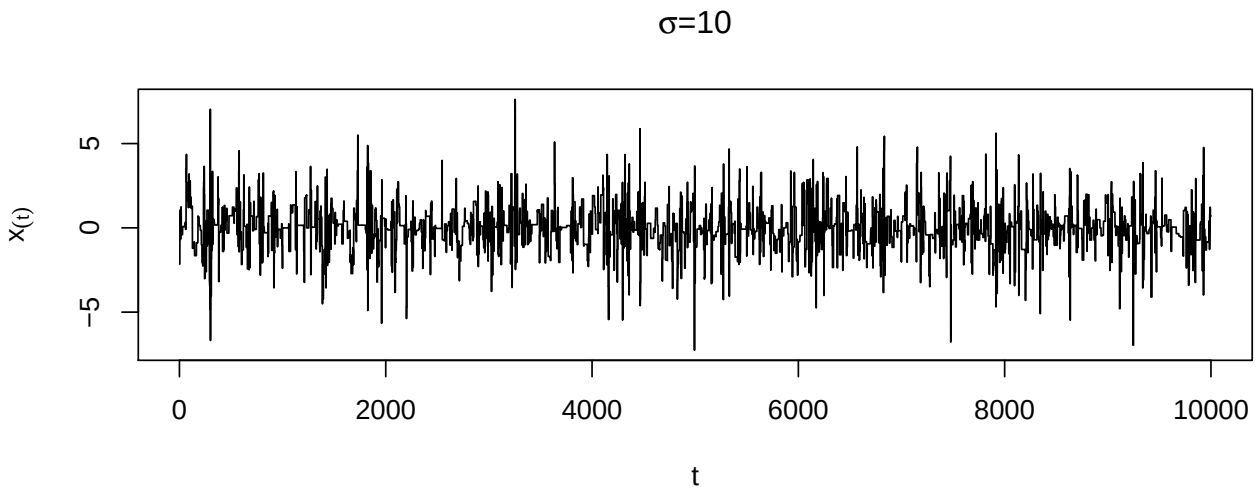
- Assume we want to construct a random walk MH sampler to generate a sample of 10,000 observations from the Laplace distribution,

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

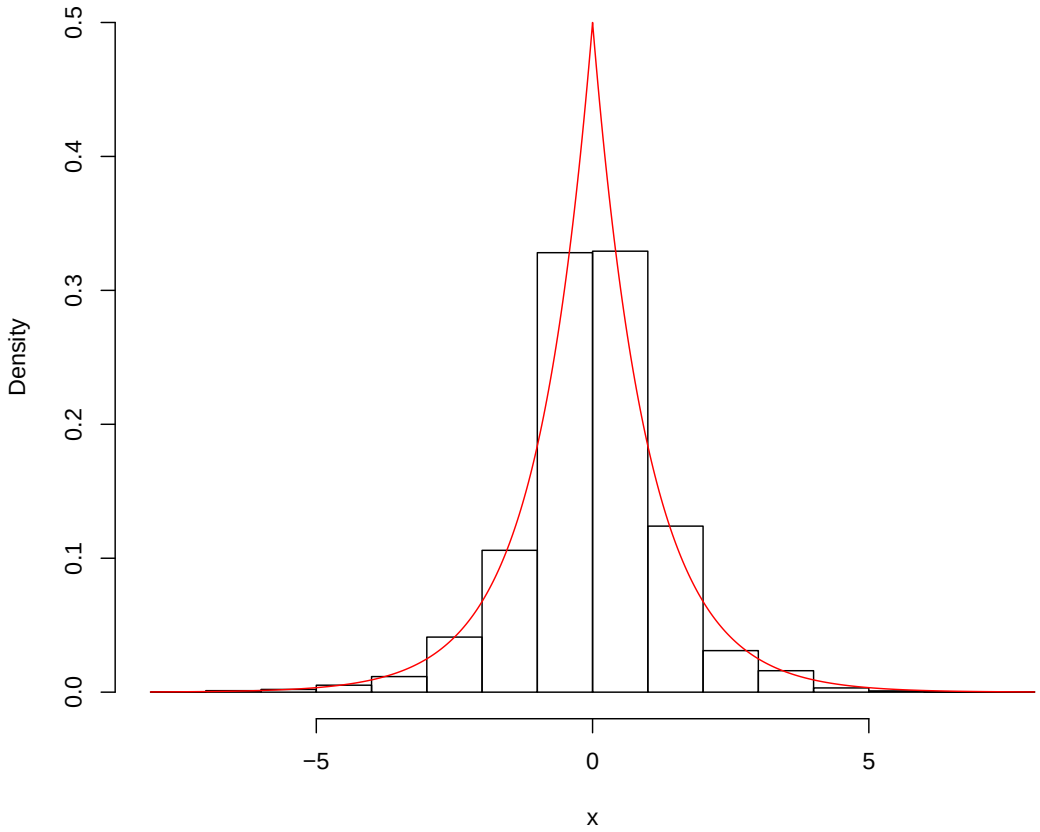
- We use a random-walk chain with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to generate proposals $X^* = x^{(t)} + \epsilon$.

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

Simulation of multidimensional distributions

- Thus far we have treated $X^{(t)}$ with little regard to its dimensionality. The Gibbs sampler is specifically adapted for multidimensional target distributions.
- The goal is to construct a Markov chain whose stationary distribution – or some marginalization thereof – equals the target distribution f.
- The Gibbs sampler does this by sequentially sampling from univariate conditional distributions, which are often available in closed form.

Overview

Notations and basic assumption

- Recall $X = (X_1, \ldots, X_p)^T$, and denote

$$X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)^T.$$

- Suppose that the univariate conditional density of $X_i \mid X_{-i} = x_{-i}$, denoted as $f(x_i \mid x_{-i})$, is easily sampled for $i = 1, \ldots, p$.
- A general Gibbs sampling procedure can be described as follows.

Basic Gibbs sampler

1. Select starting values $x^{(0)}$, and set $t = 0$.
2. Generate, in turn,

$$
\begin{aligned}
X_1^{(t+1)} \mid \cdot &\sim f(x_1 \mid x_2^{(t)}, \ldots, x_p^{(t)}) \\
X_2^{(t+1)} \mid \cdot &\sim f(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\
&\vdots \\
X_{p-1}^{(t+1)} \mid \cdot &\sim f(x_{p-1} \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-2}^{(t+1)}, x_p^{(t)}) \\
X_p^{(t+1)} \mid \cdot &\sim f(x_p \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)})
\end{aligned}
$$

   where $\mid \cdot$ denotes conditioning on the most recent updates to all other elements of X.
3. Increment t and go to step 2.

Bayesian inference

- The Gibbs sampler is particularly useful for Bayesian applications when the goal is to make inference based on the posterior distribution of multiple parameters.
- Bayesian inference is based on the posterior distribution $f(\theta \mid y) = cf(\theta)L(\theta \mid y)$, where c is an unknown constant. When the requisite univariate conditional densities are easily sampled, the Gibbs sampler can be applied and does not require evaluation of the constant c.
- In this case the i-th step in a cycle of the Gibbs sampler at iteration t is given by draws from

$$\theta_i^{(t+1)} \mid (\theta_{-i}^{(t)}, y) \sim f(\theta_i \mid \theta_{-i}^{(t)}, y),$$

where $\theta_{-i}^{(t)} = (\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_p^{(t)})$.

Example

- Let $Y_1, \ldots, Y_n$ iid from $\mathcal{N}(\mu, h^{-1})$, where $h = 1/\sigma^2$ is the precision.
- Assume the priors $\mu \sim \mathcal{N}(\mu_0, h_0^{-1})$ and $h \sim G(\alpha_0/2, \delta_0/2)$, where G denotes the Gamma distribution,

$$f(h) \propto h^{\alpha_0/2-1} \exp(-\delta_0 h/2).$$

- The posterior density is

$$f(\mu, h \mid y) \propto \underbrace{h^{n/2} \exp\left(-\frac{h}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)}_{L(\mu,h|y)} \times \underbrace{\exp\left(-\frac{h_0}{2}(\mu-\mu_0)^2\right)}_{f(\mu)} \underbrace{h^{\alpha_0/2-1}\exp\left(-\frac{\delta_0 h}{2}\right)}_{f(h)}.$$

Example (continued)

- We can compute the conditional posterior distribution of h as

$$\begin{aligned} f(h \mid \mu, y) &\propto L(\mu, h \mid y) f(h) \\ &\propto h^{(\alpha_0+n)/2-1} \exp\left\{-\frac{h}{2}\left(\delta_0 + \sum_{i=1}^{n}(y_i-\mu)^2\right)\right\} \\ &\sim G\left(\frac{\alpha_0+n}{2}, \frac{\delta_0+\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right) \end{aligned}$$

- Now

$$f(\mu \mid h, y) \propto L(\mu, h \mid y) f(\mu) \propto \exp\left\{-\frac{h_0+hn}{2}\left(\mu - \frac{h_0\mu_0+hn\overline{y}}{h_0+hn}\right)^2\right\}$$

$$\sim \mathcal{N}\left(\frac{h_0\mu_0+hn\overline{y}}{h_0+hn}, (h_0+hn)^{-1}\right)$$

Properties of the Gibbs sampler

- Clearly the chain produced by a Gibbs sampler is Markov.
- Under rather mild conditions, it can be shown that the stationary distribution of the Gibbs sampler chain is f.
- It also follows that the limiting marginal distribution of $X_i^{(t)}$ equals the univariate marginalization of the target distribution along the i-th coordinate.
- As with the MH algorithm, we can use realizations from the chain to estimate the expectation of any function of X.

Relation with the MH algorithm

- The Gibbs sampler can be seen as a special case of the MH algorithm, where
  - The proposal distribution varies over time;
  - The proposal is always accepted.
- Each Gibbs cycle consists of p MH steps.
- To see this, consider for simplicity and without loss of generality the case $p = 2$.

Relation with the MH algorithm (continued)

- In the first step of the Gibbs cycle we propose $X^* = (X_1^*, x_2^{(t)})$ given $x^{(t)} = (x_1^{(t)}, x_2^{(t)})$ from the proposal distribution

$$g_1(x^* \mid x^{(t)}) = \begin{cases} f(x_1^* \mid x_2^{(t)}) & \text{if } x_2^* = x_2^{(t)}, \\ 0 & \text{otherwise.} \end{cases}$$

- The MH ratio is

$$R(x^{(t)}, x^*) = \frac{f(x^*) g_1(x^{(t)} \mid x^*)}{f(x^{(t)}) g_1(x^* \mid x^{(t)})} = \frac{f(x^*) f(x^{(t)}) / f(x_2^{(t)})}{f(x^{(t)}) f(x^*) / f(x_2^{(t)})} = 1.$$

So, the proposal is accepted and we set $x_1^{(t+1)} = x_1^*$.

Relation with the MH algorithm (continued)

- Similarly, in the second step, we propose $X^* = (x_1^{(t+1)}, x_2^*)$ given $x^{(t+\frac{1}{2})} = (x_1^{(t+1)}, x_2^{(t)})$ from the proposal distribution

$$g_2(x^* \mid x^{(t+\frac{1}{2})}) = \begin{cases} f(x_2^* \mid x_1^{(t+1)}) & \text{if } x_1^* = x_1^{(t+1)}, \\ 0 & \text{otherwise.} \end{cases}$$

- The MH ratio is

$$\begin{aligned} R(x^{(t+\frac{1}{2})}, x^*) &= \frac{f(x^*) g_2(x^{(t+\frac{1}{2})} \mid x^*)}{f(x^{(t)}) g_2(x^* \mid x^{(t+\frac{1}{2})})} \\ &= \frac{f(x^*) f(x^{(t+\frac{1}{2})}) / f(x_1^{(t+1)})}{f(x^{(t+\frac{1}{2})}) f(x^*) / f(x_1^{(t+1)})} = 1. \end{aligned}$$

Again, the proposal is accepted and we set $x_2^{(t+1)} = x_2^*$.

Overview

Variants of the Gibbs sampler

- The "Gibbs sampler" is actually a generic name for a rich family of very adaptable algorithms.
- We will now see some strategies that have been developed to improve the performance of the general algorithm just described.

Update ordering

- The ordering of updates made to the components of X in the basic Gibbs sampler can change from one cycle to the next. This is called random scan Gibbs sampling.
- Randomly ordering each cycle can be effective when parameters are highly correlated. The random scan Gibbs sampling approach can yield faster convergence rates than the deterministic update ordering.
- In practice, it may be a good strategy to try both deterministic and random scan Gibbs sampling when parameters are found to be highly correlated.

Blocking

- Another modification to the Gibbs sampler is called blocking or grouping. In the Gibbs algorithm it is not necessary to treat each element of X individually.
- In the basic Gibbs sampler with $p = 4$, for example, it would be allowable for each cycle to proceed with the following sequence of updates:

$$
\begin{aligned}
X_1^{(t+1)} \mid \cdot &\sim f\left(x_1 \mid x_2^{(t)}, x_3^{(t)}, x_4^{(t)}\right) \\
X_2^{(t+1)}, X_3^{(t+1)} \mid \cdot &\sim f\left(x_2, x_3 \mid x_1^{(t+1)}, x_4^{(t)}\right) \\
X_4^{(t+1)} \mid \cdot &\sim f\left(x_4 \mid x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}\right)
\end{aligned}
$$

- Blocking is typically useful when elements of X are correlated, with the algorithm constructed so that more correlated elements are sampled together in one block.

Hybrid Gibbs sampling

- For many problems the conditional distributions for one or more elements of X are not easily sampled.
- In this case, a hybrid MCMC algorithm can be developed where at a given step in the Gibbs sampler, the MH algorithm is used to sample from the appropriate conditional distribution.

Hybrid Gibbs sampling: example

For example, for $p = 5$, a hybrid MCMC algorithm might proceed with the following sequence of updates:

1. Update $X_1^{(t+1)} \mid (x_2^{(t)}, x_3^{(t)}, x_4^{(t)}, x_5^{(t)})$ with a Gibbs step because this conditional distribution is easily sampled.
2. Update $(X_2^{(t+1)}, X_3^{(t+1)}) \mid (x_1^{(t+1)}, x_4^{(t)}, x_5^{(t)})$ with a MH step because this joint conditional distribution is difficult to sample from. Here, blocking X_2 and X_3 might be recommended because these elements are highly correlated.
3. Update $X_4^{(t+1)} \mid (x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_5^{(t)})$ with a step from a random walk chain because this conditional distribution is not easily sampled.
4. Update $X_5^{(t+1)} \mid (x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_4^{(t+1)})$ with a Gibbs step.

Overview

Implementation issues

- All of the MCMC methods described so far have the correct limiting stationary distribution. In practice, however, it is necessary to determine when the chain has run sufficiently long so that the output adequately represents the target distribution and can be used reliably for estimation.
- Unfortunately, MCMC methods can sometimes be quite slow to converge, requiring extremely long runs, especially if the dimensionality of X is large.

Questions

In this section, we examine questions about the long-run behavior of the chain, such as:

- Has the chain run long enough?
- Has the chain traversed all portions of the region of support of f?
- Are the sampled values approximate draws from f?
- How shall the chain output be used to produce estimates and assess their precision?

Overview

Necessity of diagnostic tools

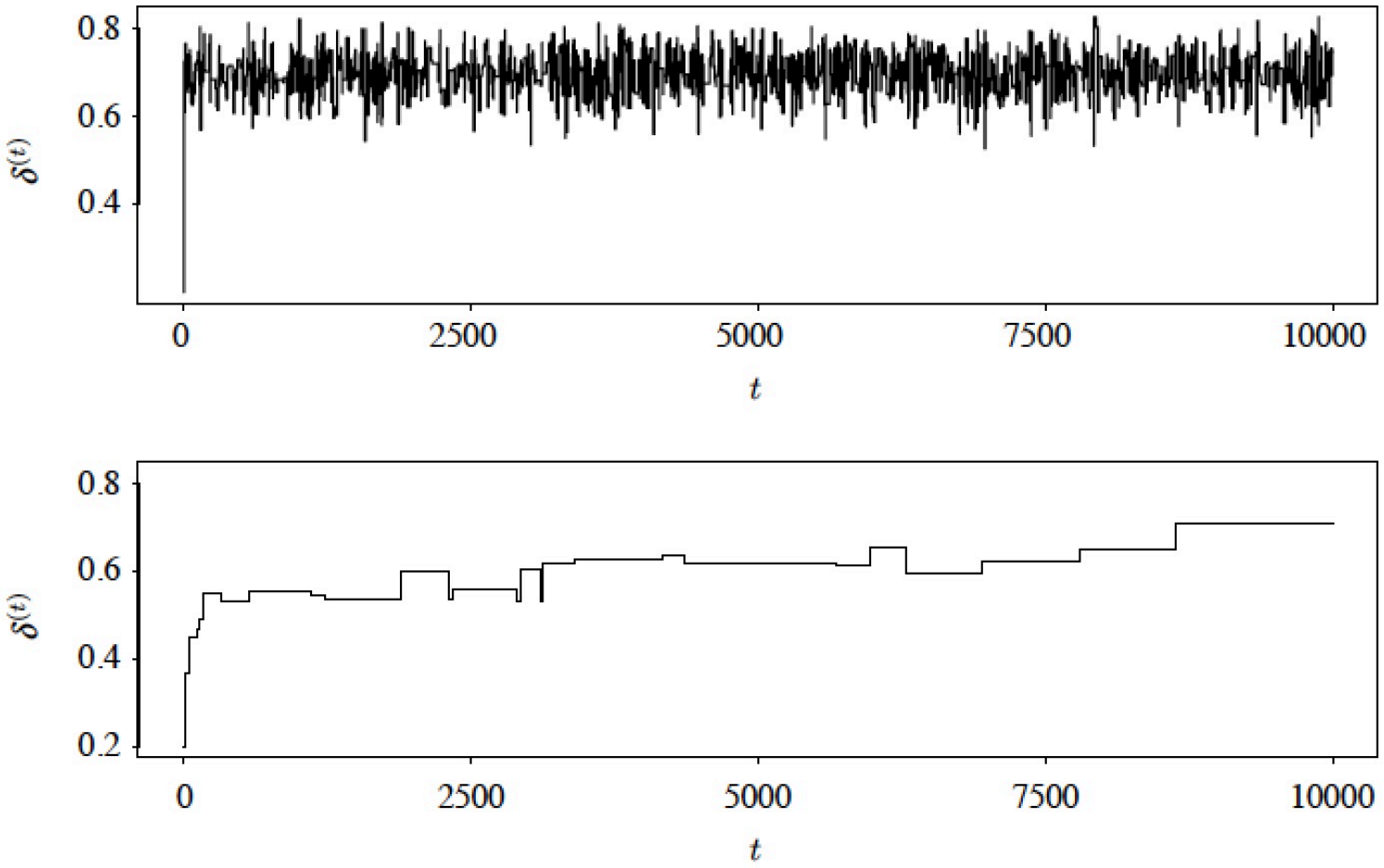
- Two main issues:
  1. Mixing: how quickly the chain forgets its starting value, how quickly the chain fully explores the support of the target distribution, how far apart in a sequence observations need to be before they can be considered to be approximately independent.
  2. Convergence: Has the chain approximately reached its stationary distribution?
- There is substantial overlap between the goals of diagnosing convergence to the stationary distribution and investigating the mixing properties of the chain. The same diagnostics can be used to investigate both mixing and convergence.
- It is recommended to use a variety of diagnostic techniques.

Simple Graphical Diagnostics

Two main graphics:

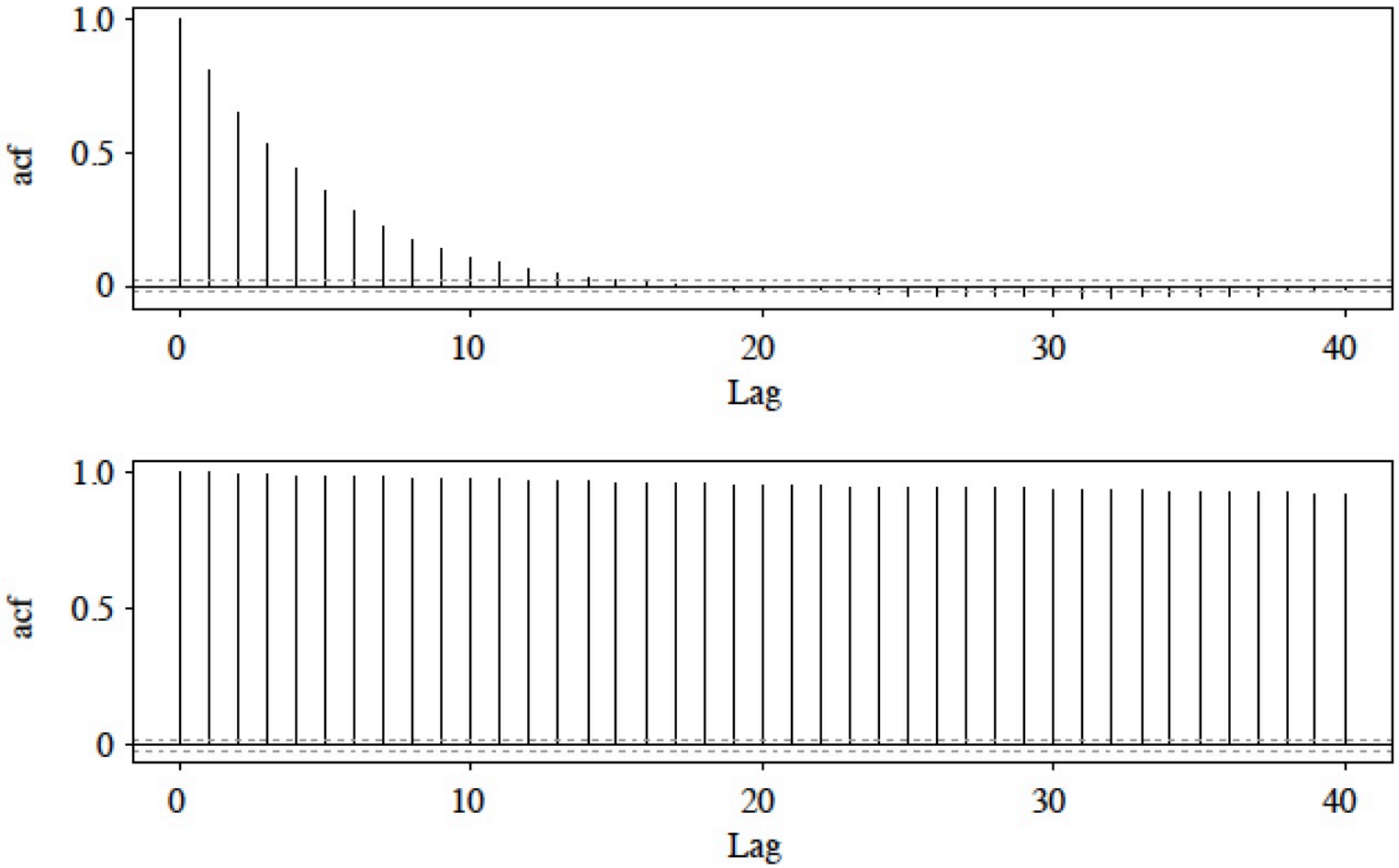
1. The sample path is a plot of the iteration number t versus the realizations of $X^{(t)}$. If a chain is mixing poorly, it will remain at or near the same value for many iterations. A chain that is mixing well will quickly move away from its starting value and the sample path will wiggle about vigorously in the region supported by f.
2. The autocorrelation plot summarizes the correlation in the sequence of $X^{(t)}$ at different iteration lags. The autocorrelation at lag k is the correlation between iterates that are k iterations apart. A chain that has poor mixing properties will exhibit slow decay of the autocorrelation as the lag between iterations increases.

Example of sample paths



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

Example of autocorrelation plot



Autocorrelation function plots for the independence chain with proposal densities Beta(1,1) (top) and Beta(2,10) (bottom).

Burn-in

- Key considerations in the diagnosis of convergence are the burn-in period and run length.
- Recall that it is only in the limit that an MCMC algorithm yields $X^{(t)} \sim f$.
- For any implementation, the iterates will not have exactly the correct marginal distribution, and the dependence on the initial point (or distribution) from which the chain was started may remain strong.
- To reduce the severity of this problem, the first D values from the chain may be discarded as a burn-in period. Typically D is fixed to a few hundred or thousand values.
- Burn-in is not needed if we start the chain in a region of high density.

Choice of proposal

- Mixing is strongly affected by features of the proposal distribution, especially its spread. Further, advice on desirable features of a proposal distribution depends on the type of MCMC algorithm employed.
- For a general MH chain such as an independence chain, it seems intuitively clear that we wish the proposal distribution g to approximate the target distribution f very well, which in turn suggests that a very high rate of accepting proposals is desirable.
- Although we would like g to resemble f, the tail behavior of g is more important than its resemblance to f in regions of high density. In particular, if f/g is bounded, the convergence of the Markov chain to its stationary distribution is faster overall. Thus, it is wiser to aim for a proposal distribution that is somewhat more diffuse than f.

Effective Sample Size

- If MCMC realizations are highly correlated, then the information gained from each iteration of the MCMC algorithm will be much less than suggested by the run length.
- The effective sample size is the size of an iid sample that would contain the same quantity of information.
- To estimate the effective sample size, we first compute the autocorrelation time defined as

$$\tau = 1 + 2\sum_{k=1}^{\infty} \rho(k),$$

where $\rho(k)$ is the autocorrelation with lag k.

- A common approach to estimate τ is to truncate the summation when $\widehat{\rho}(k) < 0.1$.
- Then the effective sample size for an MCMC run with L iterations after burn-in can be estimated using $L/\widehat{\tau}$.

Comparing chains

- Effective sample size can be used to compare the efficiency of competing MCMC samplers for a given problem.
- For a fixed number of iterations, an MCMC algorithm with a larger effective sample size is likely to converge more quickly.
- For example, we may be interested in the gains achieved from blocking in a Gibbs sampler. If the blocked Gibbs sampler has a much higher effective sample size than the unblocked version, this suggests that the blocking has improved the efficiency of the MCMC algorithm.

Overview

Standard one-number summary statistics

- Standard one-number summary statistics such as means and variances are commonly desired.
- The most commonly used estimator is based on an empirical average. Discard the burn-in; then calculate the desired statistic by taking

$$\widehat{\mu} = \frac{1}{L} \sum_{t=D}^{D+L-1} h(X^{(t)})$$

as the estimator of $\mu = \mathbb{E}\{h(X)\}$, where L denotes the length of each chain after discarding D burn-in iterates. This estimator is consistent even though the $X^{(t)}$ are serially correlated.

Simulation standard error

- The Monte Carlo, or simulation, standard error (sse) of an estimator is also of interest. This is an estimate of the variability in the estimator if the MCMC algorithm were to be run repeatedly.
- The naive estimate of the standard error for an estimator like μ is the sample standard deviation of the L realizations after burn-in divided by $\sqrt{L}$.
- However, MCMC realizations are typically positively correlated, so this procedure can underestimate the standard error.
- A simple estimator of the standard error is the batch means method.

Batch means method

- By examining the empirical autocorrelations, determine a lag k_0 such that the autocorrelation is small enough to be neglected, e.g., $\widehat{\rho}(k_0) \leq 0.05$.
- Then divide the L observations after burn-in into $L/k_0 = B$ batches.
- Let $\widehat{\mu}_b$ be the mean of $h(X^{(t)})$ in batch b. The sample variance of the means is

$$S^2 = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\mu}_b - \widehat{\mu})^2$$

and the estimated simulation standard error is

$$\widehat{sse}(\widehat{\mu}) = \sqrt{S^2/B}.$$