Simulation - Lectures 8 - MCMC: Gibbs

Lecture version: Monday 9th March, 2020, 12:22

Robert Davies

Part A Simulation and Statistical Programming

Hilary Term 2020

Outline

Metropolis Hastings

Gibbs sampler

Example: Gaussian mixture model

- Let $p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$, where π is a vector of mixture proportions, and p_k is a normal density with mean μ_k and variance σ_k^2
- Here, suppose $\pi_1 = 0.5$ with $\mu_1 = 3, \mu_2 = 5$ and $\sigma_1 = \sigma_2 = 1$
- MH algorithm with target pdf p and proposal transition pdf

$$q(y|x) = \begin{cases} 1 & \text{for } y \in [x - 1/2, x + 1/2] \\ 0 & \text{otherwise} \end{cases}$$

- Acceptance probability

$$\alpha(y|x) = \min\left(1, \frac{p(y)q(x|y)}{p(x)q(y|x)}\right) = \min\left(1, p(y)/p(x)\right)$$

Code

```
set.seed(9110)
p <- function(x) {
    0.5 * dnorm(x, mean = 3) + 0.5 * dnorm(x, mean = 5)
}
n <- 10000
x <- numeric(n)
x[1] <- 4
for(t in 1:(n - 1)) {
    yt <- x[t] + (runif(1) - 0.5)
    if (runif(1) < (p(yt) / p(x[t]))) {
        x[t + 1] <- yt
    } else {
        x[t + 1] <- x[t]
    }
}
```

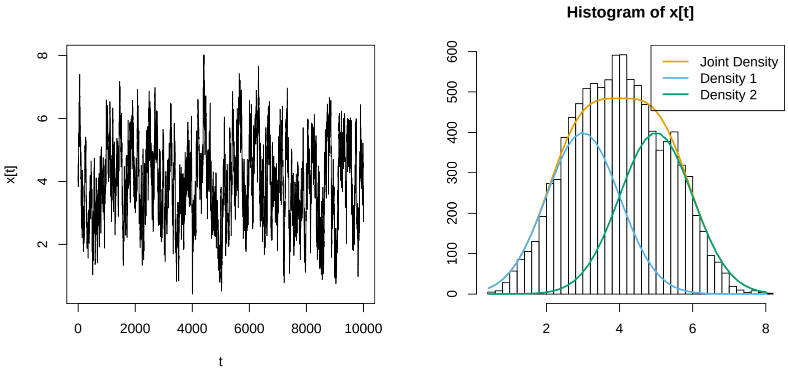


Figure: $\mu_1 = 3, \mu_2 = 5$

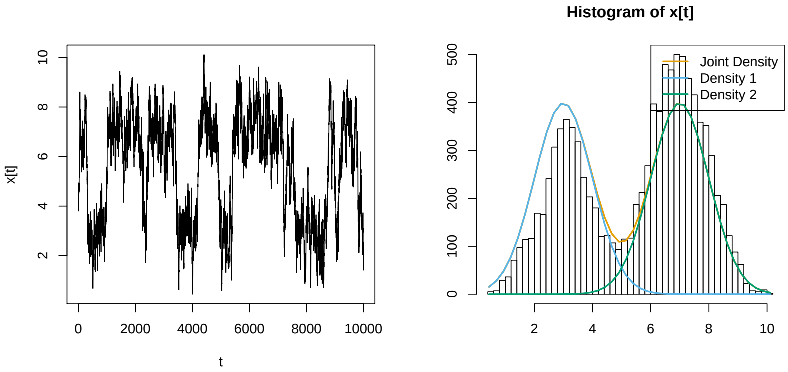


Figure: $\mu_1 = 3, \mu_2 = 7$

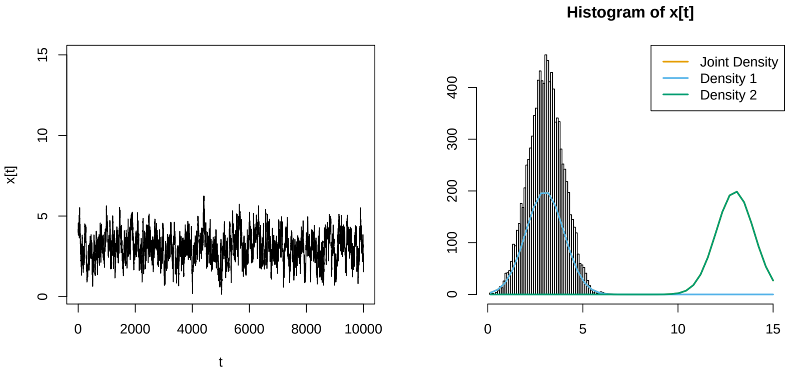


Figure: $\mu_1 = 3, \mu_2 = 13$

Bivariate normal example

- Consider a Metropolis algorithm for simulating two-dimensional samples $Z_t = (X_t, Y_t)$ from a bivariate Normal distribution with mean $\mu = (0, 0)$ and covariance

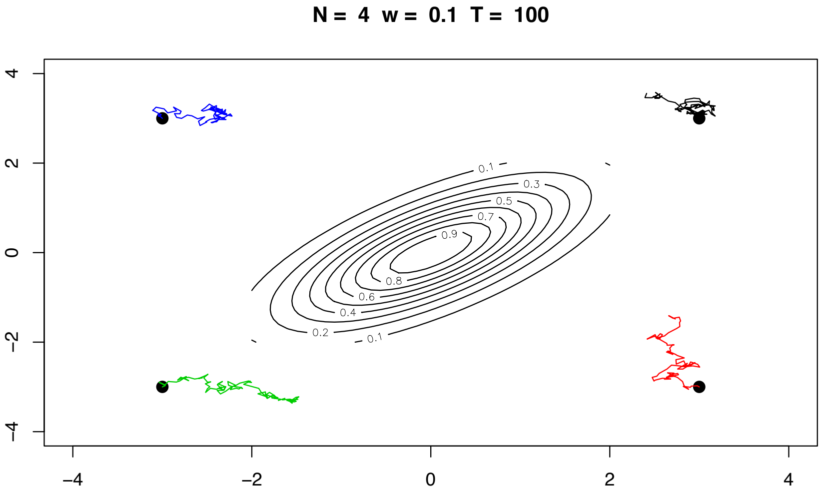
$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

- Using a proposal distribution that considers only the current value for that dimension

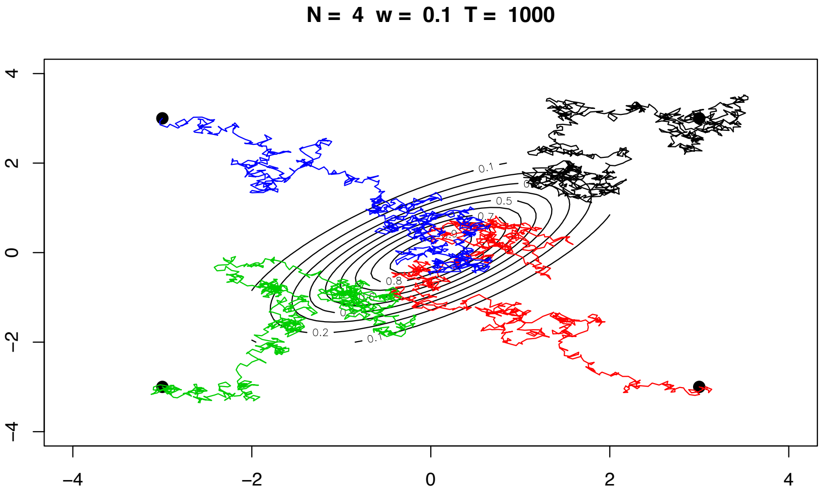
$$X_{t+1}|X_t = x, Y_t \sim U(x - w, x + w) \quad i \in \{1, 2\}$$

- The performance of the algorithm can be controlled by setting w
  - If w is small then we propose smaller moves and the chain will move slowly
  - If w is large then we propose larger moves and may accept only a few moves
- There is an 'art' to implementing MCMC that involves choice of good proposal distributions.

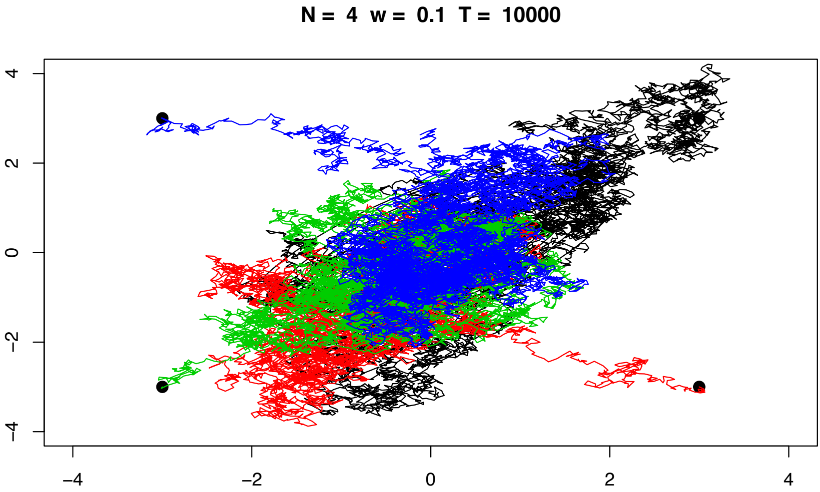
Example



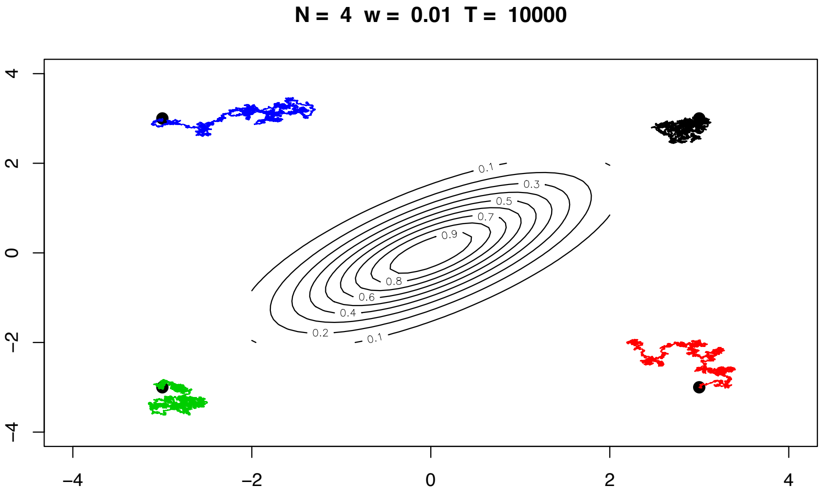
Example



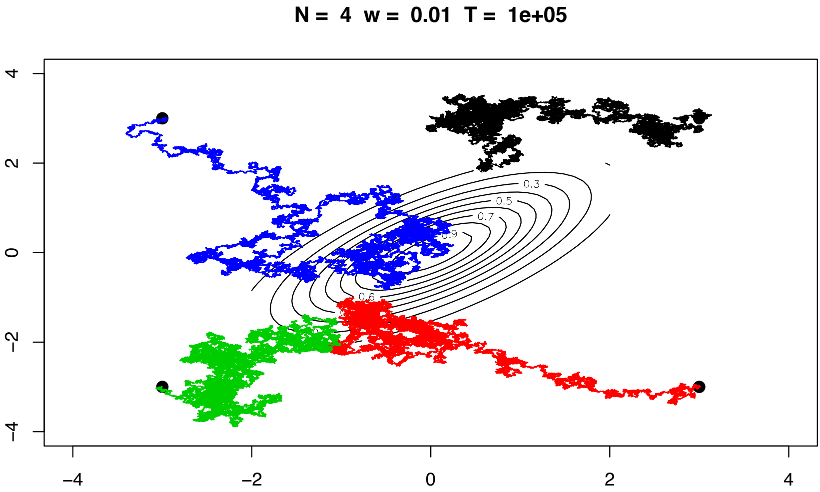
Example



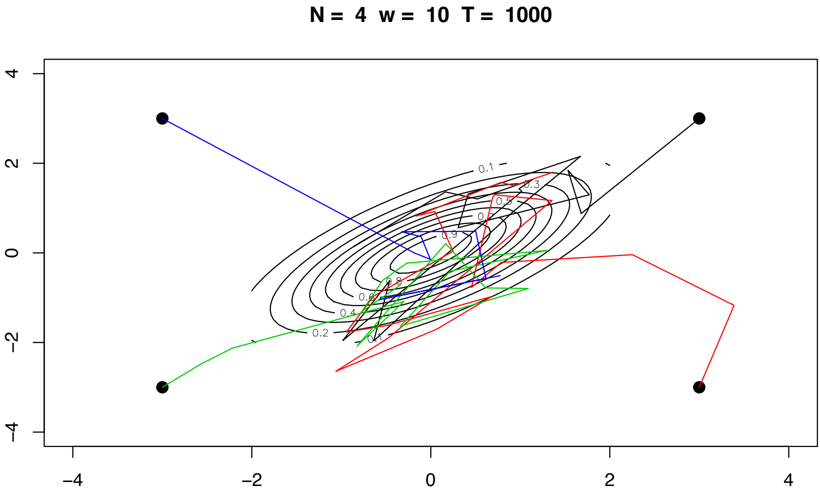
Example



Example



Example



MCMC: Practical aspects

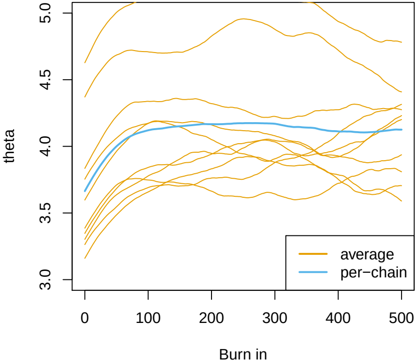
- The MCMC chain does not start from the stationary distribution, so $\mathbb{E}[\phi(X_t)] \neq \mathbb{E}[\phi(X)]$ and the difference can be significant for small t
- X_t converges in distribution to p as $t \to \infty$
- Common practice is to discard the b first values of the Markov chain $X_0, \ldots, X_{b-1}$, where we assume that X_b is approximately distributed from p
- We use the estimator

$$\frac{1}{n-b} \sum_{t=b}^{n-1} \phi(X_t)$$

- The initial $X_0, \ldots, X_{b-1}$ is called the burn-in period of the MCMC chain.

Burn in example

- Consider the first example of this lecture (GMM) for $\theta = \mathbb{E}[X]$ for $\mu_1 = 3, \mu_2 = 5$.
- Let $X_0 = -5$. This is an unlikely start $p(-5) = 2 \times 10^{-15}$
- Note that in real world, choosing good start points is non-trivial
- Below, estimates of θ wrt burn-in across chains given fixed $n = 2000$



Outline

Metropolis Hastings

Gibbs sampler

MCMC: Gibbs Sampler

- The Gibbs sampling algorithm is a Markov Chain Monte Carlo (MCMC) algorithm to simulate a Markov chain with a given stationary distribution
- Unlike with MCMC Metropolis Hastings, where we use a proposal and acceptance / rejection, in Gibbs sampling, we make use of conditional distributions
- In the proof that follows we will assume a discrete distribution, but it follows for continuous distributions (unproved)
- Note we will use notation $x_{-j} = (x_1, ..., x_{j-1}, x_{j+1}, \ldots, x_d\}$ to indicate that the entries of x minutes the j^{th} entry
- Further note we will use notation like x_j^t to refer to the j^{th} entry of x^t, where x^t is the $t + 1^{st}$ realization / sample of the Markov chain

Gibbs Sampler algorithm

Gibbs Sampler algorithm

1. Let p be the target pmf with conditional $p(x_j^t|x_{-j}^t)$ for x with d dimensions
2. Either set $X^0 = x^0$, or draw X^0 from some initial distribution
3. For $t = 1, 2, \ldots, n-1$:
   3.1 Assume $X^{t-1} = x^{t-1}$, and set $X^t = x^{t-1}$
   3.2 Choose a permutation S of $\{1, ..., d\}$
   3.3 For $j \in S$:
      3.3.1 Sample X_j^t from $p(x_j^t|x_{-j}^t)$

Gibbs Sampler about

- The Gibbs sample defines a Markov chain with transition matrix P, such that, for $x, y \in \Omega$ with $x_i = y_i \ \forall i \neq j$, with $x_j = a$ and $y_j = b$, that

$$P_{x,y} = p(x_j = b | x_{-j}) = \frac{P(x_1, \ldots, x_j = b, \ldots)}{\sum_{a^*} P(x_1, ..., x_j = a^*, ...)}$$

Gibbs Sampler theorem and proof

Theorem

The transition matrix P of the Markov chain generated by the Gibbs sampler is reversible with respect to p and therefore admits p as a stationary distribution.

- Proof: We check detailed balance. For $x \neq y \in \Omega$ with $x_i = y_i \ \forall i \neq j$, with $x_j = a$ and $y_j = b$, that

$$
\begin{aligned}
p(x)P_{x,y} &= p(x)p(x_j = b | x_{-j}) \\
&= p(x_1 ..., x_j = a, ...)\frac{P(x_1, \ldots, x_j = b, \ldots)}{\sum_{b^*} P(x_1, ..., x_j = b^*, ...)} \\
&= p(x_1 ..., x_j = b, ...)\frac{P(x_1, \ldots, x_j = a, \ldots)}{\sum_{a^*} P(x_1, ..., x_j = a^*, ...)} \\
&= p(y)p(x_j = a | x_{-j}) \\
&= p(y)P_{y,x}
\end{aligned}
$$

Gibbs sampling and Metropolis-Hastings

- Recall that in MH we have an acceptance probability given by

$$\alpha(y|x) = \min\left\{1, \frac{p(y)q(x|y)}{p(x)q(y|x)}\right\}$$

- Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions given by the conditional $p(x_j = b|x_{-j})$. We then get that for $x_j = a$ and $y_j = b$ with $x_i = y_i$ otherwise, that

$$\begin{aligned}\frac{p(y)q(x|y)}{p(x)q(y|x)} &= \frac{p(y)p(x_j = a|x_{-j})}{p(x)p(x_j = b|x_{-j})} \\ &= \frac{p(\ldots, x_j = b, \ldots)p(\ldots, x_j = a, \ldots)/\sum_{a^* p(\ldots, x_j = a^*, \ldots)}}{p(\ldots, x_j = a, \ldots)p(\ldots, x_j = b, \ldots)/\sum_{b^* p(\ldots, x_j = b^*, \ldots)}} \\ &= 1\end{aligned}$$

- That is, every proposed move is accepted

GMM example, revisited

- Return to Gaussian Mixture Model example from beginning of lecture
- Let $p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$, where π is a vector of mixture proportions, and p_k is a normal density with mean μ_k and variance $\sigma_k^2 = 1$
- Consider a latent (hidden) variable Z of whether we are in the first or second distribution, so that $P(Z = k) = \pi_k$ and

$$X|Z = k \sim N(\mu_k, \sigma_k)$$

$$Z|X = x \sim P(Z = k|X = x) = \frac{f_k(x|\mu_k, \sigma_k)\pi_k}{\sum_{j=1}^{K} f_j(x|\mu_j, \sigma_j)\pi_j}$$

- We can jointly estimate the distribution of (X, Z), and then just consider the X that we want

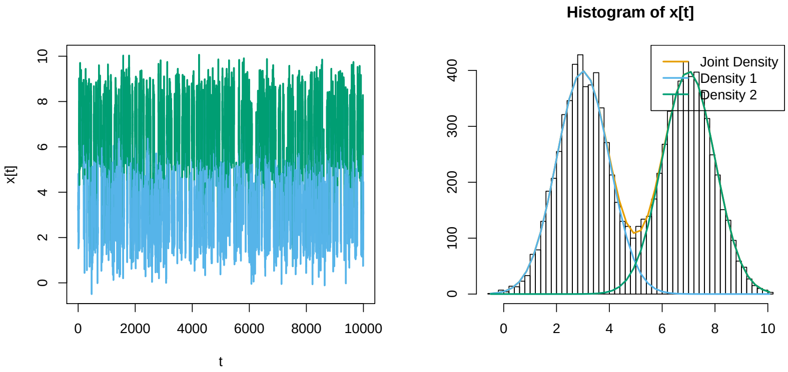


Figure: $\mu_1 = 3, \mu_2 = 7$

Code

```
set.seed(9110)
n <- 10000
pi <- c(0.5, 0.5); mus <- c(3, 7)
x <- numeric(n); z <- numeric(n)
x[1] <- 4; z[1] <- 1
for(t in 1:(n - 1)) {
    x[t + 1] <- x[t]; z[t + 1] <- z[t]
    for(s in sample(1:2)) {
        if (s == 1) {    ## sample z
            u <- dnorm(x[t + 1], mean = mus[1]) * pi[1]
            l <- dnorm(x[t + 1], mean = mus[2]) * pi[2]
            choose1 <- runif(1) < (u / (u + l))
            z[t + 1] <- c(2, 1)[as.integer(choose1) + 1]
        } else { ## sample x
            x[t + 1] <- rnorm(1, mean = mus[z[t + 1]], sd = 1)
        }
    }
}
```

Bivariate normal example revisited

- Consider from earlier this lecture about the bivariate normal to generate samples $Z_t = (X_t, Y_t)$ from a bivariate Normal distribution with mean $\mu = (0, 0)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

- In the general case for a bivariate normal with means μ_x, μ_y, variances σ_x^2, σ_y^2 and correlation ρ, once can derive (not shown here) the marginal distribution of $f(y|x)$ as

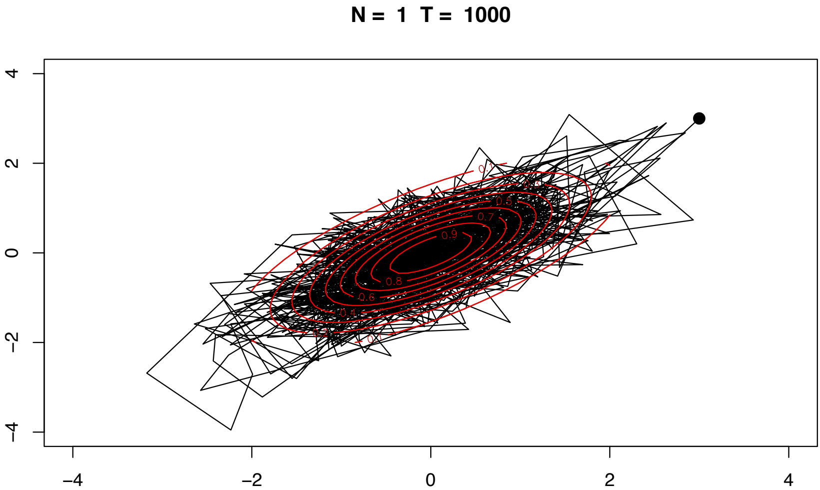
$$X_{t+1}|Y_t = y \sim N(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y), \sigma_x^2(1 - \rho^2))$$

- Which in our case is

$$X_{t+1}|Y_t = y \sim N(0.7\frac{1}{\sqrt{2}}y, (1 - 0.7^2))$$

- In other words, we initialize (X_0, Y_0) at some value, and conditionally sample X and Y given their marginal distributions

Bivariate normal example revisited, example



Recap

- When using MCMC in practice, one must carefully consider tuning parameters and the proposal distribution to avoid insufficnient mixing / exploration of the space, as well as burn-in
- Gibbs sampling is a particular form of MCMC where we use conditional probabilities instead of a proposal
- It is akin to Metropolis Hastings with an acceptance probability of 1