Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Gibbs sampler
- Metropolis-Hastings algorithm
- Accuracy of MCMC approximations

Recap of MCMC basics

See blackboard

Gibbs sampler

The Gibbs sampler

- Assume $X \in \mathbb{R}^p$ and divide X in ***k* components** $X = (X_1, X_2, \ldots, X_k)$

- Denote the **conditional density of the *i*-th component X_i given all the other components $X_{-i} = (X_j)_{j \neq i}$ by π_i**:

$$\pi_i(x_i \mid x_{-i}) \propto \pi(x)$$

  where $\propto$ means up to a term which does not contain x_i

- The π_is are called **full conditionals**

- Since $\pi_i(x_i|x_{-i}) \propto \pi(x)$, we can identify $\pi_i(x_i|x_{-i})$ by inspecting $\pi(x)$

The Gibbs sampler

The Gibbs sampler depends on a **visiting schedule** $i_t \in \{1, 2, \ldots, k\}$ and iterates the following steps for $t = 1, 2, \ldots$

$$X_{i_t}^t \sim \pi_{i_t}(x_{i_t} \mid X_{-i_t}^{t-1}), \quad X_{-i_t}^t = X_{-i_t}^{t-1}$$

- Leave all components of X^{t-1} unchanged except the one that is visited, and update the visited component according to the conditional distribution
- The visiting schedule can be either deterministic or random

  In order that the chain can reach all sets, we have to visit each possible component infinitely often

The Gibbs sampler

Gibbs sampler (with fix visiting schedule)

1. Simulate $X^0 = (X_1^0, X_2^0, \ldots, X_k^0)$
2. For $t = 1, 2, \ldots,$ simulate
    1. $X_1^t \sim \pi_1(x_1 | X_2^{t-1}, \ldots, X_k^{t-1})$
    2. $X_2^t \sim \pi_2(x_2 | X_1^t, X_3^{t-1}, \ldots, X_k^{t-1})$

       ...

    i. $X_i^t \sim \pi_i(x_i | X_1^t, \ldots, X_{i-1}^t, X_{i+1}^{t-1}, \ldots, X_k^{t-1})$

       ...

    k. $X_k^t \sim \pi_k(x_k | X_1^t, \ldots, X_{k-1}^t)$

See R example and blackboard

Clicker question

Metropolis-Hastings algorithm

Reversibility

- A distribution π is called **reversible** for the transition kernel P if

$$\int_A \pi(x)P(x,B)dx = \int_B \pi(x)P(x,A)dx \quad \forall A,B$$

  I.e., if $X^t \sim \pi$, then

$$\mathbb{P}(X^t \in A, X^{t+1} \in B) = \mathbb{P}(X^{t+1} \in A, X^t \in B) \quad \forall A,B$$

- A **reversible distribution π is also invariant** (choose B as the whole space $\mathbb{R}^p$)

- If $P(x,.)$ has the density $p(x,y)$ for any x, then reversibility is equivalent to

$$\pi(x)p(x,y) = \pi(y)p(y,x) \quad \forall x,y$$

  See blackboard

Construction of the Metropolis-Hastings algorithm

- The **Metropolis-Hastings algorithm** generates a chain which has π **as reversible distribution**
- Can we simply choose one of the two values $p(x, y)$ and $p(y, x)$ arbitrarily and then determine the other one by the above equation?

  No, since then in general $\int p(x, y)dy = 1$ does not hold true for all x

Construction of the Metropolis-Hastings algorithm

Solution to the above problem:

1. Choose an arbitrary transition density q
2. Select from the two possible solutions

$$p(x,y) = q(x,y),\ p(y,x) = \frac{\pi(x)q(x,y)}{\pi(y)}$$

and

$$p(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)},\ p(y,x) = q(y,x)$$

the one which satisfies both $p(x,y) \leq q(x,y)$ and $p(y,x) \leq q(y,x)$ for any $x \neq y$

Construction of the Metropolis-Hastings algorithm

- This can be written in the **compact form**

$$p(x,y) = q(x,y)a(x,y)$$

where

$$a(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$$

- It follows that $\int p(x,y)dy \leq \int q(x,y)dy = 1$ for any x, and the **remaining mass is put on the "diagonal"**:

$$p(x,x) = 1 - \int p(x,y)dy$$

- In summary, the **transition kernel** can be written as

$$P(x,A) = \int_A p(x,y)dy + 1_A(x)\left(1 - \int p(x,y)dy\right)$$

The Metropolis-Hastings algorithm

Metropolis-Hastings (MH) algorithm

1. Simulate X^0
2. For $t = 1, 2, \ldots,$
   - 2a. Generate $Y^t \sim q(X^{t-1}, x)dx$ and $U^t \sim$ uniform(0,1), independently from each other and independently of previously generated variables
   - 2b. Set
   $$X^t = \begin{cases} Y^t & \text{if} \quad U^t \leq a(X^{t-1}, Y^t) \\ X^{t-1} & \text{else} \end{cases}$$

Comments on the Metropolis-Hastings algorithm

- The MH algorithm is similar to rejection sampling, but when a proposed value Y_t is rejected, we keep the current value, in accordance with the definition $P(x, x) = 1 - \int p(x, y)dy$
- q is called the **proposal distribution** and a is called the **acceptance probability**

See blackboard

Clicker question

The random walk Metropolis (RWM) algorithm

- An often used choice of $q(X^{t-1}, .)$ is a normal density with mean X^{t-1} and an arbitrary covariance matrix Σ, i.e.,

$$\boldsymbol{Y^t \sim \mathcal{N}(X^{t-1}, \Sigma)}$$

- In this case and in general **if $q(x, y) = q(y, x)$, the acceptance probability simplifies to**

$$\min\left(1, \frac{\pi(y)}{\pi(x)}\right)$$

- **Interpretation**: if the probability of the proposed value $\pi(y)$ is greater than the probability of the current value $\pi(x)$, one always accepts the proposed value, otherwise only with some probability < 1

See R example and blackboard

Combination of Gibbs and Metropolis-Hastings algorithm

- One can also **combine different proposal densities for different components** (e.g. Gibbs steps with random walk Metropolis steps)

Accuracy of MCMC approximations

Accuracy of MCMC approximations

- Determining how reliable MCMC approximations are is not always easy

There are the following two difficulties:

1. There is a **bias**:

$$\mathbb{E}(\bar{h}_{N,r}) \neq \int h(x)\pi(x)dx$$

2. Since **successive values X^t are dependent, the variance is more complicated**:

$$\mathrm{Var}\left(\bar{h}_{N,r}\right) = \frac{1}{(N-r)^2}\left(\sum_{t=r+1}^{N} \mathrm{Var}(h(X^t)) + 2\sum_{t=r+1}^{N}\sum_{s=1}^{N-t} \mathrm{Cov}(h(X^t), h(X^{t+s}))\right)$$

Accuracy of MCMC approximations

A pragmatic way to deal with these complications:

1. **Look at trace plots** of $h(X^t)$ or of components X_i^t versus t and choose r such that the series “looks stationary” for $t \geq r$

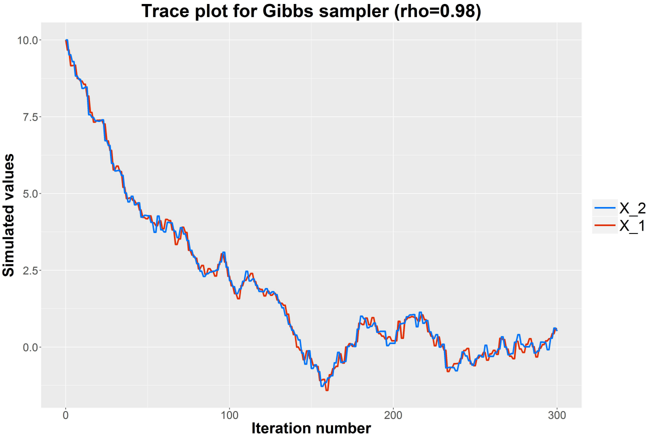
2. Assume that $X^t \sim \pi$ for $t \geq r$ so that
   - There is no bias
   - The **covariances** $\mathrm{Cov}(h(X^t), h(X^{t+s}))$ depend only on s and **can be estimated** by

$$\frac{1}{N-r}\sum_{t=r+1}^{N-s}(h(X^t)-\bar{h}_{N,r})(h(X^{t+s})-\bar{h}_{N,r})$$

- The number of replicates N should then be large enough that these estimated covariances are close to zero for most lags s

Example of trace plot

Gibbs sampler for bivariate normal distribution



$\Rightarrow$ Guess burn-in time of approx. $r = 200$