Markov Chain Monte Carlo methods

Stat 580: Statistical Computing

- Theme: Black - White
- Printable version

General information

- A series of random variables $\{Y_i, i = 1, \ldots, n\}$ has the Markov property if the conditional distribution of Y_i, given all the the previous observations $Y_1, \ldots, Y_{i-1}$ depends only on Y_{i-1}.
- In these two slides, we assume that the state space is countable for ease of exposition.
- If Y_i has a discrete sample space (state space), this is the same as
$$P(Y_i = k|Y_1, Y_2, \ldots, Y_{i-1}) = P(Y_i = k|Y_{i-1} = j) = P_{jk},$$
where P_{jk} is the probability that the variable "jumps" from state j to state k, known as transition probability.
- We only focus on the homogeneous case when P_{jk} does not depend on i.
- Transition matrix $\boldsymbol{P} = \{P_{jk}\}$
- $\{Y_1, \ldots\}$ is called a Markov chain, characterized by $\boldsymbol{P}$.

General information

- If $\boldsymbol{P}$ possesses the following properties:
  1. Irreducibility: the Markov chain is not made up of smaller cycles
  2. Aperiodicity: $P(X_j = k|X_0 = k) > 0$ and $P(X_{j+1} = k|X_0 = k) > 0$
  3. Recurrence: the Markov chain returns to its original state with probability 1.
- Then a stationary distribution (limiting equilibrium distribution) exists, denoted as π or f.

Weather example

In the summer, each day in Ames is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7, whereas a rainy day is followed by a sunny day with probability 0.4. It rains on Monday (M). Make weather forecasts for Tuesday (T), Wednesday (W), and Thursday (H), using a homogeneous Markov chain model.

- Forecast for Tuesday (one-step transition)

$$P(\text{T sunny} \mid \text{M rainy}) = 0.4$$
$$P(\text{T rainy} \mid \text{M rainy}) = 0.6$$

- Forecast for Wednesday (two-step transition): e.g.,

$$\begin{aligned} P(\text{W sunny} \mid \text{M rainy}) &= P(\text{W sunny} \mid \text{T rainy, M rainy})P(\text{T rainy} \mid \text{M rainy}) \\ &\quad + P(\text{W sunny} \mid \text{T sunny, M rainy})P(\text{T sunny} \mid \text{M rainy}) \\ &= (0.4)(0.6) + (0.7)(0.4) = 0.52 \end{aligned}$$

Weather example

- let "sunny" and "rainy" be state 1 and 2 respectively.
- Using matrix notation,
  - the transition matrix

$$\boldsymbol{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

  - Forecast for Tuesday (one-step transition): second row of $\boldsymbol{P}$
    - can be regarded as $(0, 1)\boldsymbol{P}$
  - Forecast for Wednesday (two-step transition): second row of $\boldsymbol{P}^2$
    - can be regarded as $(0, 1)\boldsymbol{P}^2$

Weather example

- Forecast for 14 days later:

$$\boldsymbol{P}^{14} = \begin{pmatrix} 0.5714286 & 0.4285714 \\ 0.5714285 & 0.4285715 \end{pmatrix}$$

- Forecast for 30 days later:

$$\boldsymbol{P}^{30} \simeq \begin{pmatrix} 0.5714286 & 0.4285714 \\ 0.5714286 & 0.4285714 \end{pmatrix} = \begin{pmatrix} 4/7 & 3/7 \\ 4/7 & 3/7 \end{pmatrix}$$

- Note that

$$(p, 1-p) \begin{pmatrix} 4/7 & 3/7 \\ 4/7 & 3/7 \end{pmatrix} = (4/7, 3/7)$$

  - no matter what the starting distribution $(p, 1-p)$ is, the distribution after long time converge to $\pi = (4/7, 3/7)$ (stationary distribution)

Markov Chain Monte Carlo

- often, the goal is to generate a series of observations whose stationary distribution π (or f) is proportional to the unnormalized posterior distribution (mainly for Bayesian methods)

- these observations are correlated, so we may need to "skip every m observations" to get an (approximate) *iid* sample

- also, it will take some time for the chain to reach its equilibrium state, so throw away the first e.g. 10000 observations (known as burn-in).

Gibbs sampling

- goal: sample from joint distribution
- applicable when the conditional distributions of variables (given others) can be easily constructed (and sampled) from
- basic form for three components: for $i = 1, \ldots, n$, do
  - Generate X_i from $f(x|y = Y_{i-1}, z = Z_{i-1})$.
  - Generate Y_i from $f(y|x = X_i, z = Z_{i-1})$.
  - Generate Z_i from $f(z|x = X_i, y = Y_i)$.
- then the triple (X_i, Y_i, Z_i) forms a Markov chain whose stationary distribution is the joint distribution of X, Y, Z
- note that any or all of the three components (X, Y, Z) may be multivariate

Example

Suppose we are trying to simulate from $f(x, y) = \binom{n}{x} y^{x+n-1}(1-y)^{n-x+\beta-1}$, where x is an integer from 0 to n, and $y \in [0, 1]$.

- $f(x|y) \propto \binom{n}{x} y^x (1-y)^{n-x}$, which is $\text{Binomial}(n, y)$.
- $f(y|x) \propto y^{x+\alpha-1}(1-y)^{n-x+\beta-1}$, which is $\text{Beta}(x+\alpha, n+\beta)$.
- To use Gibbs sampling, for $i = 1, \ldots, n$, do
    1. generate Y_i from $f(y|x = X_{i-1})$
    2. generate X_i from $f(x|y = Y_i)$

Metropolis-Hastings algorithm

- a dependent version of the rejection algorithm (depend on previous draw)
- target stationary distribution $f(x)$
- need a proposal distribution with density $q(y|x)$
- The algorithm: Choose X_0. The algorithm generates X_i, for $i = 1, \ldots,$, as follows.

1. Generate Y_i from $q(y|X = x_{i-1})$ and U_i from $\mathrm{Unif}(0, 1)$.
2. Evaluate $r(X_{i-1}, Y_i)$ where
$$r(x, y) = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}.$$
3. If $U_i \leq r(X_{i-1}, Y_i)$, then set $X_i = Y_i$. If not, then set $X_i = X_{i-1}$.

Metropolis-Hastings algorithm

- If we only know the density $f(x)$ up to a constant c (i.e., $f(x) = cp(x)$), we can still use Metropolis-Hastings.
- Common choices for $q(y|x)$ is $\mathcal{N}(x, b^2)$ for some $b > 0$. In this case, q is symmetric, i.e., $q(y|x) = q(x|y)$, and r simplifies to
$$r(x, y) = \min\left\{\frac{f(y_i)}{f(x_{i-1})}, 1\right\}.$$
- X_i can be a vector of random variables, or other object like a tree structured "data point".

Example

Let $f(x) = \frac{1}{\pi(1+x^2)}$ be the distribution we are trying to generate from (Cauchy distribution).

- Take $q(y|x)$ as $\mathcal{N}(x, b^2)$. So
$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}, 1\right\} = \min\left\{\frac{1+x^2}{1+y^2}, 1\right\}$$
- The algorithm is:
  1. Draw $Y_i \sim \mathcal{N}(X_{i-1}, b^2)$.
  2. Set
$$X_i = \begin{cases} Y_i & \text{with probability } r(X_{i-1}, Y_i), \\ X_{i-1} & \text{with probability } 1 - r(X_{i-1}, Y_i). \end{cases}$$

Example

- If b is small (e.g., $b = 0.1$), the chain takes small steps and does not "explore" much of the sample space.

- If b is too small, the chain will be highly correlated.

- If b is large (e.g., $b = 10$), the proposals (Y_i) are often far in the tails, making r small and hence we often reject the proposal.

- For this example $b = 1$ is about the right choice.

Why does it work?

- Let $p(x, y)$ be the probability of making a transition from x to y.
- Our goal is to show that $f(x)$ in the algorithm satisfies $f(x) = \int f(y)p(y, x)dy$, which implies $f(x)$ is a stationary distribution of the chain.
- Consider two points x and y. Either:
$$f(x)q(y|x) < f(y)q(x|y),$$
or
$$f(x)q(y|x) > f(y)q(x|y).$$
We ignore ties, which happen with probability 0 for continuous distributions.

Why does it work?

- Without loss of generality, assume that $f(x)q(y|x) > f(y)q(x|y)$. Then,

  $$r(x, y) = \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}$$

  and

  $$r(y, x) = 1$$

  since we have assumed already (WLOG) that the numerator is larger than the denominator.

Why does it work?

- $p(x, y)$ is the probability of jumping from x to y, which happens in two stages. We first generate y and then we must accept it. Mathematically,

$$\begin{aligned} p(x, y) &= \underbrace{q(y|x)}_{\text{probability of generating } y} \cdot \underbrace{r(x, y)}_{\text{probability of accepting}} \\ &= q(y|x) \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)} \\ &= \frac{f(y)}{f(x)} q(x|y), \end{aligned}$$

so

$$f(x)p(x, y) = f(y)q(x|y).$$

Why does it work?

- Similarly,

$$p(y, x) = q(x|y)r(y, x) = q(x|y),$$

and therefore,

$$\begin{aligned} & & f(x)p(x, y) &= f(y)p(y, x) \\ &\Rightarrow & \int f(x)p(x, y)dy &= \int f(y)p(y, x)dy \\ &\Rightarrow & f(x)\int p(x, y)dy &= \int f(y)p(y, x)dy \\ &\Rightarrow & f(x) &= \int f(y)p(y, x)dy \end{aligned}$$