LEC 07

MCMC

I. Markov Chains

A Markov chain is a discrete-time stochastic process $\theta_1, \theta_2, \dots, \theta_t, \dots$ that satisfies the Markov property

$$p(\theta_t \mid \theta_{t-1}, \dots, \theta_1) = p(\theta_t \mid \theta_{t-1}) \quad \forall t.$$

state of the chain

$$\to \theta_{t-1} \to \theta_t \to$$

A Markov chain is fully defined by

- $\pi_0(\theta)$, the initial distribution.
- $T(\theta_{t-1}, \theta_t) = p(\theta_t \mid \theta_{t-1})$, the transition kernel.

Basic limit theorem

$\pi_n(\theta) = \int \pi_{n-1}(\theta')\, T(\theta', \theta)\, d\theta'$

Let $\pi_n = \pi_0 \underbrace{T T \ldots T}_{n \text{ times}}$ be the probability over the states after n iterations.

Let $\pi_*(\theta)$ be the stationary distribution of the Markov chain with kernel T, i.e. such that $\pi_* T = \pi_*$.

Thm $\lim_{n \to \infty} \pi_n(\theta) = \pi_*(\theta) \quad \forall \theta$

"π_n converges to π_*"

Assumptions:

- π_* exists
- The chain is irreducible, i.e. if any state can be reached from any other state with positive probability in a finite number of steps.

- The chain is aperiodic.

A state θ is periodic with period k if the number of steps to return to θ is always divisible by $k \geq 2$.

A Markov chain is aperiodic if none of its states is periodic with $k \geq 2$.

Time reversibility

A Markov chain is time reversible if

$$(\theta_0, \theta_1, \dots, \theta_n) \overset{D}{=} (\theta_n, \theta_{n-1}, \dots, \theta_0)$$

"equal in distribution"

$$\Rightarrow (\theta_0, \theta_1) \overset{D}{=} (\theta_1, \theta_0)$$

The sequence of states moving forward is equal in distribution to the sequence of states moving backward.

and $(\theta_0) \overset{D}{=} (\theta_1)$

$$\Downarrow$$

$$\pi_0 = \pi_1$$
$$= \pi_0 T \quad \Rightarrow \pi_0 \text{ is the stationary distribution}$$

Also $(\theta_0, \theta_1) \overset{?}{=} (\theta_1, \theta_0)$

$$\iff P(\theta_0 = i, \theta_1 = j) = P(\theta_1 = i, \theta_0 = j) \quad \forall i, j$$

$$\iff P(\theta_0 = i) P(\theta_1 = j \mid \theta_0 = i)$$
$$= P(\theta_0 = j) P(\theta_1 = i \mid \theta_0 = j)$$

$$\iff \underbrace{\pi_0(i) T(i,j) = \pi_0(j) T(j,i)}_{\text{Local balanced equations}}$$

If the local balanced equation hold for π and T, then π is the stationary distribution governed by T.

II. Metropolis - Hastings (MH)

MH sampling works by defining a Markov chain $\theta_1, \theta_2, \ldots$ whose stationary distribution is the target $\pi(\theta)$.

candidate

current state

① $\theta' \sim q(\theta' | \theta_t)$

proposal distribution

② $\alpha(\theta' | \theta_t) = \min \left\{ \frac{\pi(\theta')}{\pi(\theta_t)} \frac{q(\theta_t | \theta')}{q(\theta' | \theta_t)}, 1 \right\}$

acceptance ratio

③ $\theta_{t+1} = \begin{cases} \theta' & \text{with probability } \alpha(\theta' | \theta_t) \\ \theta_t & \text{otherwise} \end{cases}$

$u \sim U[0;1]$

if $u < \alpha$, accept

⚠ MH only requires the ratio of the target density $\dfrac{\pi(\theta')}{\pi(\theta_t)}$

$\Rightarrow$ We can use an unnormalised density $\pi^*(\theta) = \frac{1}{Z}\pi(\theta)$ with an unknown normalizer $\frac{1}{Z}$.

Random walk MH (= "Metropolis algorithm")

$$\theta' = \theta_t + \varepsilon$$

from a symmetric and centred distribution g (e.g. $\mathcal{N}$)

In this case,

$$q(\theta'|\theta_t) = g(\varepsilon)$$

$$= \; q(\theta_t|\theta') = g(-\varepsilon) = g(\varepsilon)$$

$$\alpha(\theta'|\theta_t) = \min\left\{ \frac{\pi(\theta')}{\pi(\theta_t)} \frac{\cancel{q(\theta_t|\theta')}}{\cancel{q(\theta'|\theta_t)}}, 1 \right\}$$

Why does it work?

Thm: The MH acceptance probability $\alpha(\theta'|\theta_t)$ for the proposal $q(\theta'|\theta_t)$ and target $\pi(\theta)$ defines a Markov chain with the transition kernel

$$T(\theta_t, \theta') = \alpha(\theta'|\theta_t)\, q(\theta'|\theta_t).$$

This chain satisfies the detailed balanced condition.

Proof:

$$\pi(\theta_t)\, T(\theta_t, \theta') = \pi(\theta_t)\, \alpha(\theta'|\theta_t)\, q(\theta'|\theta_t)$$

$$= \min\{\pi(\theta')\, q(\theta_t|\theta'), \pi(\theta_t)\, q(\theta'|\theta_t)\}$$

$$= \min\{\pi(\theta_t)\, q(\theta'|\theta_t), \pi(\theta')\, q(\theta_t|\theta')\}$$

$$= \pi(\theta')\, \alpha(\theta_t|\theta')\, q(\theta_t|\theta')$$

$$= \pi(\theta_t)\, T(\theta'|\theta_t) \qquad \square$$

III. Bayesian inference with MCMC

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)}$$

$$= \int p(x|\theta)\, p(\theta)\, d\theta$$

Difficult to evaluate!

$$= \frac{1}{Z(x)}\, p(x|\theta)\, p(\theta) = \pi^*(\theta)$$

$\Rightarrow$ Use MCMC for posterior sampling

$$\log \pi^*(\theta) = \log p(x|\theta) + \log p(\theta) + c$$

Diagnostics:

Gelman-Rubin diagnostic

* Convergence checking based on m chains from widely diverse starting points.

Assuming θ_{ij} $(i=1..n\ ,\ j=1...m)$

- Between-chains variance

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\theta_{:j} - \theta_{::}\right)^2$$ (variance of the mean)

where $$\theta_{:j} = \frac{1}{n} \sum_{i=1}^{n} \theta_{ij}$$

$$\theta_{::} = \frac{1}{m} \sum_{j=1}^{m} \theta_{:j}$$

- Within-chain variance

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2$$ (mean of chain variances)

where $$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\theta_{ij} - \theta_{:j}\right)^2$$

$$\Rightarrow \mathbb{V}[\theta] = \frac{n-1}{n} W + \frac{1}{n} B$$

overestimates π's variance

$$\Rightarrow \hat{R} = \sqrt{\frac{\mathbb{V}[\theta]}{W}} \geq 1$$

$\hat{R}$ approaches 1 as n increases

underestimate π's variance

→ Stop when $\hat{R} < 1.1$

* Effective sample size

$$\rho_k = \frac{\mathbb{E}\left[(\theta_i - \mu)(\theta_{i+k} - \mu)\right]}{\sigma^2}$$

over the time index i

Number of samples

$$\text{ESS} = \frac{M}{\underbrace{1 + 2\sum_{k=1}^{\infty} \rho_k}_{= \int \text{autocorrelation}}}$$

$\Rightarrow$ Uncorrelated samples $\rightarrow \rho_k = 0$

and $\text{ESS} = M$

$\Rightarrow$ Correlated samples $\rightarrow$ small ESS