Lecture Notes for STAT546: Computational Statistics

—Lecture 2: Monte Carlo

Faming Liang

Purdue University

August 21, 2024

Why do we need MCMC?

For some problems, the exact sampler is not available or very expensive, we have to resort to MCMC for simulating some approximate samples for inference.

Some convergence theorems (Tierney, 1994)

Theorem 1

Suppose $\{X_n\}$ *is an irreducible, aperiodic Markov chain with transition kernel* P *and invariant distribution* π*. Then*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \to 0$$

for π*-almost all* x*.*

Theorem 2

Suppose $\{X_n\}$ *is an irreducible Markov chain with transition kernel* P *and invariant distribution* π*, and let* h *be a real-valued function such that* $E_\pi|h(x)| < \infty$*. Then* $P(\bar{h}(x) \to E_\pi h(x)) = 1$*, for* π*-almost all* x*, where* $\bar{h}(x) = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$*.*

How to constrcut such a Markov chain with the desired invariant distribution $\pi(x)$?

General Metropolis-Hastings Recipe

- Start with any $\boldsymbol{x}^{(0)}$ and a proposal distribution $T(\cdot \to \cdot)$.
- Compute the ratio

$$r = \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(t)})} \frac{T(\boldsymbol{y} \to \boldsymbol{x}^{(t)})}{T(\boldsymbol{x}^{(t)} \to \boldsymbol{y})}$$

- Accept/Rejection decision:

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{y} & \text{with } p = \min(1, r) \\ \boldsymbol{x}^{(t)} & \text{otherwise} \end{cases}$$

Why Does It Work?

It satisfies the detailed balance condition:

$$\begin{aligned}
&\pi(\boldsymbol{x})T(\boldsymbol{x}\to\boldsymbol{y})\min\{1,\frac{\pi(\boldsymbol{y})T(\boldsymbol{y}\to\boldsymbol{x})}{\pi(\boldsymbol{x})T(\boldsymbol{x}\to\boldsymbol{y})}\}\\
&=\min\{\pi(\boldsymbol{x})T(\boldsymbol{x}\to\boldsymbol{y}),\pi(\boldsymbol{y})T(\boldsymbol{y}\to\boldsymbol{x})\}\\
&=\pi(\boldsymbol{y})T(\boldsymbol{y}\to\boldsymbol{x})\min\{1,\frac{\pi(\boldsymbol{x})T(\boldsymbol{x}\to\boldsymbol{y})}{\pi(\boldsymbol{y})T(\boldsymbol{y}\to\boldsymbol{x})}\}
\end{aligned}$$

Write $P(x,y)=T(\boldsymbol{x}\to\boldsymbol{y})\min\{1,\frac{\pi(\boldsymbol{y})T(\boldsymbol{y}\to\boldsymbol{x})}{\pi(\boldsymbol{x})T(\boldsymbol{x}\to\boldsymbol{y})}\}$, then we have

$$\pi(x)P(x,y)=\pi(y)P(y,x),$$

and

$$\int\pi(x)P(x,y)dx=\int\pi(y)P(y,x)dx=\pi(y),$$

i.e., $\pi(x)$ is invariant for the Markov chain.

Summary:

In the standard Markov chain theory, if the chain is irreducible, aperiodic [this is almost sure for the Metropolis-Hastings algorithm, (Tierney, 1994)], and possesses an invariant distribution, then the chain will become stationary at its traget distribution $\pi(x)$.

Therefore, if we run the chain long enough, the samples $x_{t_0+1}, \ldots x_T$ (after a burin-in period of t_0 iterations), can be regarded as approximately following the target distribution $\pi(x)$.

Proposal distribution

- Independence sampler:

$$T(X, Y) = q(Y)$$

In this case, the Metropolis ratio is

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

where $w(y) = \pi(y)/q(y)$.

- Random-walk proposal:

$$T(X, Y) = q(Y - X).$$

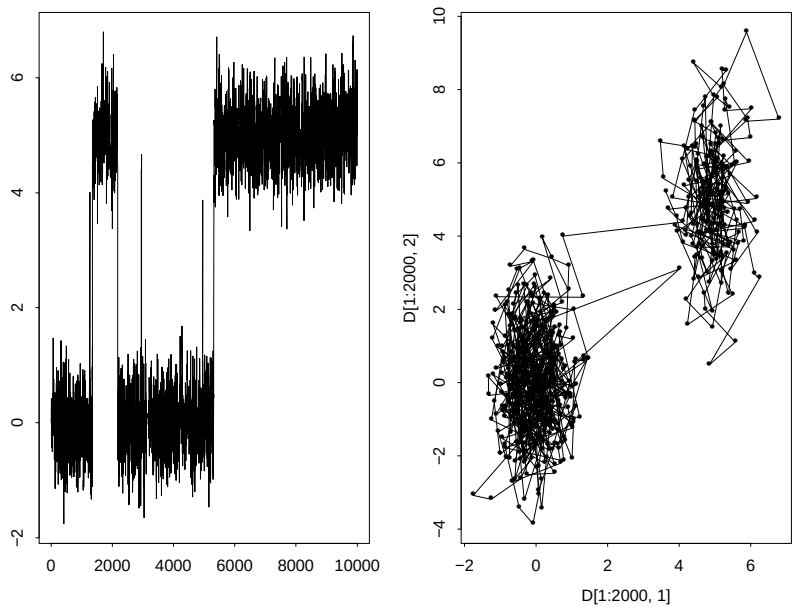
This is very popular, for example,

$$Y|X \sim N(X, \sigma^2).$$

Example: Metropolis-Hastings algorithm

- The moves are very "local".
- Tend to be trapped in a local mode.

$$\pi(x, y) \propto e^{-\frac{1}{2}\left(\frac{x^2}{0.25}+\frac{y^2}{2}\right)} + 2e^{-\frac{1}{2}\left(\frac{(x-5)^2}{0.25}+\frac{(y-5)^2}{2}\right)}$$



Gibbs sampler

- Purpose: Draw from a joint distribution $\pi(x_1, \ldots, x_k)$.
- Method: Iterative conditional sampling
  $\forall i$, Draw $x_i \sim \pi(x_i | \boldsymbol{x}_{[-i]})$.

Example: Simulating $\pi(x_1, \ldots, x_k)$.
Given the current sample $x^t = (x_1^t, \ldots, x_k^t)$,

- Draw $x_1^{t+1} \sim \pi(x_1 | x_2^t, \ldots, x_k^t)$;
- draw $x_2^{t+1} \sim \pi(x_2 | x_1^{t+1}, x_3^t, \ldots, x_k^t)$;
- $\ldots$
- draw $x_k^{t+1} \sim \pi(x_k | x_1^{t+1}, \ldots, x_{k-1}^{t+1})$.

Alternatively, $x_1, \ldots, x_k$ can be updated in a random order.

Illustrating the Gibbs sampler

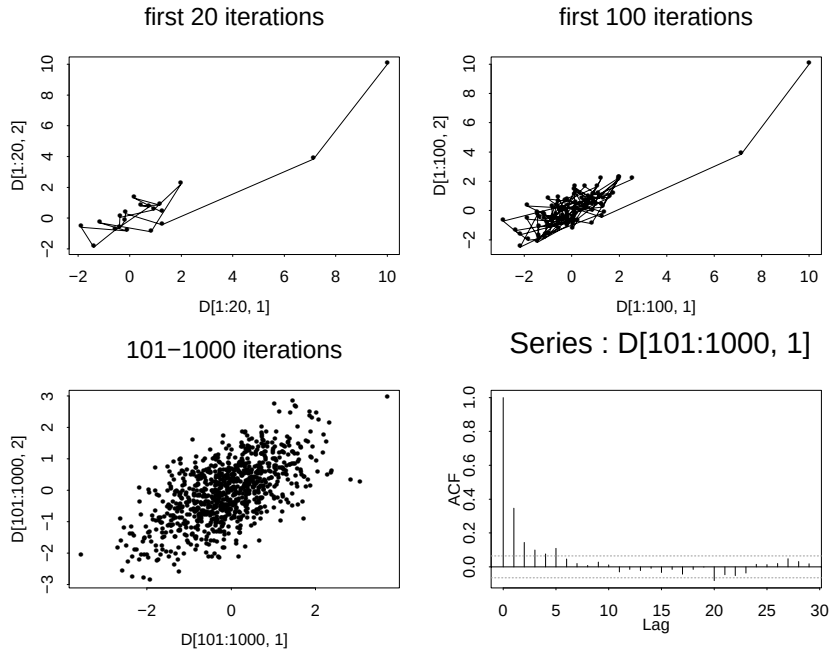
- Suppose the target distribution is

$$(x, y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

- Gibbs sampler:
  - $[X|Y = y] \sim N(\rho y, 1 - \rho^2)$
  - $[Y|X = x] \sim N(\rho x, 1 - \rho^2)$

Start form, say, $(X, Y) = (10, 10)$, we can take a look at the trajectories. We took $\rho = 0.6$.


first 20 iterations
D[1:20, 2]
D[1:20, 1]
first 100 iterations
D[1:100, 2]
D[1:100, 1]
101−1000 iterations
D[101:1000, 2]
D[101:1000, 1]
Series : D[101:1000, 1]
ACF
Lag

Why does the Gibbs sampler work?

Compute the Metropolis ratio:

$$\begin{aligned} r &= \frac{\pi(x, y^*)\pi(y|x)}{\pi(x, y)\pi(y^*|x)} \\ &= \frac{\pi(x, y^*)\pi(x, y)/\pi(x)}{\pi(x, y)\pi(x, y^*)/\pi(x)} \\ &= 1 \end{aligned}$$

A special case of the Metropolis-Hastings algorithm!

Data Augmentation

The DA algorithm can be described in the context of Bayesian analysis of incomplete data.

- X_{obs}: the observed data
- X_{mis}: the missing-data
- $X_{\text{com}} = (X_{\text{obs}}, X_{\text{mis}})$: the complete data

Suppose that the complete-data model has density $g(X_{\text{obs}}, X_{\text{mis}}|\theta)$ with the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ for some positive integer d. The objective is to make Bayesian inference with a prior distribution $p(\theta)$ for the parameter θ. Let $f(X_{\text{obs}}|\theta)$ be the observed-data model, *i.e.*,

$$f(X_{\text{obs}}|\theta) = \int_{\mathbb{X}_{\text{mis}}} g(X_{\text{obs}}, X_{\text{mis}}|\theta) dX_{\text{mis}} \qquad (\theta \in \Theta) \qquad (1)$$

For Bayesian inference about θ using MCMC methods, it requires to sampling the true or observed-data posterior

$$p(\theta|X_{\text{obs}}) \propto f(X_{\text{obs}}|\theta)p(\theta) \qquad (\theta \in \Theta) \qquad (2)$$

or more generally the joint distribution of θ and X_{mis},

$$p(\theta, X_{\text{mis}}|X_{\text{obs}}) \propto g(X_{\text{obs}}, X_{\text{mis}}|\theta)p(\theta) \qquad (\theta \in \Theta) \qquad (3)$$

Let $h(X_{\text{mis}}|\theta, X_{\text{obs}})$ be the conditional distribution of X_{mis} given θ and X_{obs}. Suppose that both $h(X_{\text{mis}}|\theta, X_{\text{obs}})$ and $p(\theta|X_{\text{obs}}, X_{\text{mis}}) \propto g(X_{\text{obs}}, X_{\text{mis}}|\theta)p(\theta)$ are easy to draw samples from. The two-step Gibbs sampler based on these two conditionals becomes a natural choice and is known as the DA algorithm. More precisely, the DA algorithm can be summarized as follows.

The DA algorithm: a two-step Gibbs sampler

Take $\theta^{(0)} \in \Theta$, *and iterate for* $t = 1, 2, ...$

I-Step. *Generate* $X_{\text{mis}}^{(t)} \sim f_{\text{mis}}(X_{\text{mis}}|\theta^{(t-1)}, X_{\text{obs}})$.

P-Step. *Generate* $\theta^{(t)} \sim p(\theta|X_{\text{obs}}, X_{\text{mis}})$.

As a two-step Gibbs sampler, DA creates two interleaving Markov chains $\{\theta^{(t)} : t = 1, 2, ...\}$ and $\{X_{\text{mis}}^{(t)} : t = 1, 2, ...\}$.

Consider a complete-data set consisting of a sample of n, $Y_1, ..., Y_n$, from the p-dimensional multivariate normal distribution $N_p(\mu, \Sigma)$ with unknown mean vector $\mu \in \mathbb{R}^p$ and $p \times p$ (positive definite) covariance matrix Σ. Each component of Y_i is either fully observed or missing for each $i = 1, ..., n$. Let $Y_{\text{obs}}^{(i)}$ denote the observed components and let $Y_{\text{mis}}^{(i)}$ the missing components of Y_i. The conditional distribution of $Y_{\text{mis}}^{(i)}$ given $Y_{\text{obs}}^{(i)}$ and (μ, Σ) is

$$Y_{\text{mis}}^{(i)} | (Y_{\text{obs}}^{(i)}, \mu, \Sigma) \sim N_{\dim(Y_{\text{mis}}^{(i)})}(\mu_{\text{mis}}^{(i)} + \Sigma_{\text{mis,obs}}^{(i)} \left[\Sigma_{\text{obs,obs}}^{(i)}\right]^{-1} (Y_{i,\text{mis}} - \mu_{\text{mis}}^{(i)})) \tag{4}$$

Suppose that for Bayesian analysis, we use the prior distribution

$$p(\mu, \Sigma) \propto |\Sigma|^{-(q+1)/2}, \tag{5}$$

where q is a known integer. With $q = p$, this prior becomes the the Jeffreys prior for Σ.

Let $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ and let $S = \sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})'$. The complete-data posterior distribution $p(\mu, \Sigma | Y_1, ..., Y_n)$ can be characterized by

$$\Sigma | Y_1, ..., Y_n \sim \frac{1}{|\Sigma|^{(n+q)/2}} \exp\left\{ -\frac{1}{2}\text{trace}\left(\Sigma^{-1}S\right)\right\}, \qquad (6)$$

that is, the inverse Wishart distribution, and

$$\mu | \Sigma, Y_1, ..., Y_n \sim N_p(\bar{Y}, \Sigma/n). \qquad (7)$$

Thus, the DA algorithm has the following I and P steps.

I-step. *For $i = 1, ..., n$, draw $Y_{\text{mis}}^{(i)}$ from (4).*

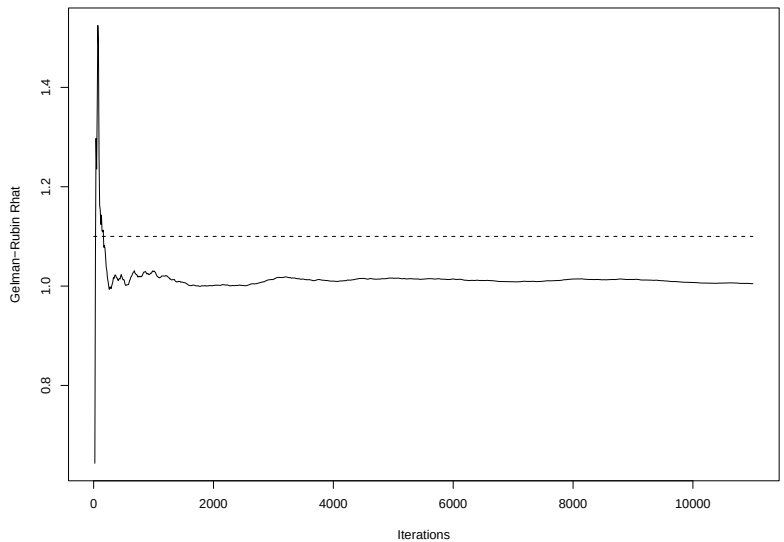
P-step. *First draw Σ from (6) given $Y_1, ..., Y_n$ and then draw μ from (7) given $Y_1, ..., Y_n$ and Σ.*

Convergence Diagnostic: Trace plot

- One intuitive and easily implemented diagnostic tool is a trace plot, which plots the sample at time t against the iteration number.
- A clear sign of non-convergence with a trace plot occurs when we observe some trending in sample space.
- For the single modal distribution, the traceplot will move snake around the mode after convergence.
- The problem with the trace plot is that it may appear we have converged, however, the chain is actually trapped in a local region rather exploring the full distribution.

Convergence Diagnostic: Gelman-Rubin Statistic

- Gelmans argues that the best way to identify non-convergence is to simulate multiple sequences with over-dispersed starting points.
- The inuition is that the behavior of all of the chains should be basically the same. Or, the variance within the chain should be the same as the variance across the chains.
- Formulas:
  - Within chain variance $W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$.
  - Between chain variance $B = \frac{n}{m-1} \sum_{i=1}^{m} (\bar{x}_i - \bar{x})^2$.
  - Estimated variance $\widehat{V} = (1 - \frac{1}{n})W + \frac{1}{n}B$.
  - Gelman-Rubin Statistic: $R = \sqrt{\widehat{V}/W}$.
- The chain is considered as converged when $R < 1.1$ or 1.2.


Gelman−Rubin Rhat
0.8
1.0
1.2
1.4
0
2000
4000
6000
8000
10000
Iterations