MT4531/5731: (Advanced) Bayesian Inference

The Gibbs sampler

Nicolò Margaritella

School of Mathematics and Statistics, University of St Andrews

–

University of St Andrews

Outline

Outline

Introduction

- The Gibbs sampler is the most popular method for sampling from joint distributions, i.e. from distributions in a multi-dimensional setting.
- It is very important for Bayesian statistics because, given data $\boldsymbol{x}$, it can be used for sampling from the joint posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ of the model parameters $\boldsymbol{\theta}$.
- In this lecture, to illustrate the main idea, we will present the Gibbs sampler for a limited number of model parameters; specifically, for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$.
- For the general mathematical representation of the algorithm please see the lecture notes on Moodle.

Outline

The Gibbs sampler (1)

- Assume a vector of parameters (random variables) $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ with distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$.
- The Gibbs sampler uses sampling from the full conditional distributions of π to sample indirectly from the full posterior distribution.
- The full conditional distributions are: $\pi(\theta_1|\theta_2, \theta_3, \boldsymbol{x})$, $\pi(\theta_2|\theta_1, \theta_3, \boldsymbol{x})$ and $\pi(\theta_3|\theta_1, \theta_2, \boldsymbol{x})$.
- The brilliance of the Gibbs sampler lies in the fact that the full conditional distributions are always univariate, and so it can be 'easy' to sample from them.
- (Within our Bayesian context, π is the posterior distribution of interest. We could drop from the notation the conditioning on the data, $\boldsymbol{x}$, as the Gibbs sampler applies to any joint distribution anyway, but we will keep this conditioning for this lecture).

The Gibbs sampler (2)

- We initially set arbitrary starting values $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \theta_3^0)$.
- Given the Markov chain is currently in state $\boldsymbol{\theta}^0$ at iteration 0 of the chain, the Gibbs sampler successively makes random drawings from the full conditional distributions, given the current values of the parameters, as follows:

$$\begin{array}{lll} \theta_1^1 & \text{is sampled from} & \pi(\theta_1 | \theta_2^0, \theta_3^0, \boldsymbol{x}) \\ \theta_2^1 & \text{is sampled from} & \pi(\theta_2 | \theta_1^1, \theta_3^0, \boldsymbol{x}) \\ \theta_3^1 & \text{is sampled from} & \pi(\theta_3 | \theta_1^1, \theta_2^1, \boldsymbol{x}) \end{array}$$

- This completes one cycle from iteration 0 to iteration 1.

The Gibbs sampler (3)

- In general, given the Markov chain is currently in state $\boldsymbol{\theta}^t$ at iteration t, the Gibbs sampler successively makes random drawings:

$$
\begin{array}{lll}
\theta_1^{t+1} & \text{is sampled from} & \pi(\theta_1|\theta_2^t, \theta_3^t, \boldsymbol{x}) \\
\theta_2^{t+1} & \text{is sampled from} & \pi(\theta_2|\theta_1^{t+1}, \theta_3^t, \boldsymbol{x}) \\
\theta_3^{t+1} & \text{is sampled from} & \pi(\theta_3|\theta_1^{t+1}, \theta_2^{t+1}, \boldsymbol{x})
\end{array}
$$

- This completes a transition from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$.
- Applying this algorithm for T iterations, produces a sequence $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^t, \ldots, \boldsymbol{\theta}^T$.
- The values $\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^B$ are discarded as burn-in, for some suitable value of B (see previous MCMC lecture), and the values $\boldsymbol{\theta}^{B+1}, \ldots, \boldsymbol{\theta}^T$ can be used to obtain Monte Carlo estimates of interest.

The Gibbs sampler (4)

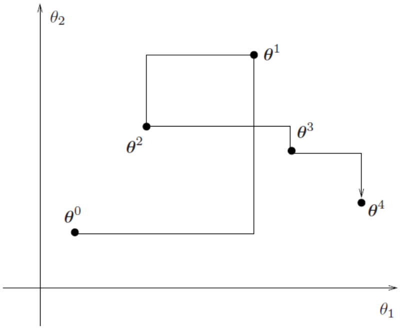
- A more general notation, used in the lecture notes, is that, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, $\pi(\theta_i|\boldsymbol{\theta}_{(i)})$ denotes the full conditional of θ_i, given the values of the other components $\boldsymbol{\theta}_{(i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$, (and the data).
- The transition kernel for going from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$ is given by

$$\mathcal{K}_G(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}) = \prod_{i=1}^{k} \pi(\theta_i^{t+1}|\theta_j^{t+1}, j < i \text{ and } \theta_j^t, j > i, \boldsymbol{x}), \quad (1)$$

and it can be proven that the stationary distribution is π.

The Gibbs sampler (5)

- In two dimensions a typical trajectory of the Gibbs sampler may look something like that:



The Gibbs sampler (6)

- `Nimble` utilises the Gibbs sampler to a great extent, in a black-box manner.
- If you want to write your own code for sampling from some multivariate posterior distribution, using the Gibbs sampler:
  - First write down the posterior distribution up to proportionality, considering $\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
  - Then, derive the full conditionals $\pi(\theta_i|\boldsymbol{\theta}_{(i)}, \boldsymbol{x})$ up to proportionality.
  - This is simple! Just notice that $\pi(\theta_i|\boldsymbol{\theta}_{(i)}, \boldsymbol{x}) \propto \pi(\boldsymbol{\theta}|\boldsymbol{x})$.
  - This is because $\boldsymbol{\theta}_{(i)}$ and $\boldsymbol{x}$ are constants as far as the full conditional distribution is concerned.
  - So, you need to look at the expression for $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ up to proportionality, and treat all factors that only depend on $\boldsymbol{\theta}_{(i)}$ and $\boldsymbol{x}$ as constants, absorbed within the proportionality sign.
  - Whatever is left on the right hand side of $\pi(\theta_i|\boldsymbol{\theta}_{(i)}, \boldsymbol{x}) \propto \pi(\boldsymbol{\theta}|\boldsymbol{x})$ gives you the full conditional distribution up to proportionality.

The Gibbs sampler (7)

- Once you have $\pi(\theta_i|\boldsymbol{\theta}_{(i)}, \boldsymbol{x})$ as a function of θ_i, up to proportionality, you need to find out how to sample from it.
- If you are lucky and recognize it as a standard distribution, you can use R functions such as 'rbeta()' or 'rnorm()' to do so.
- If it is not a standard distribution, you can use techniques we will learn in this module (e.g. Metropolis-Hastings sampling, Rejection sampling, or Importance sampling).

Outline

Example (1)

- Consider observed data $x_1, \dots, x_n$ such that, given μ and σ, each $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 are unknown.
- We specify independent priors:

$$\mu \sim N(\phi, \tau^2); \qquad \text{and} \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

Example (2)

- The posterior distribution up to proportionality is,

$$\pi(\mu, \sigma^2 | \boldsymbol{x}) \propto f(\boldsymbol{x} | \mu, \sigma^2) p(\mu) p(\sigma^2)$$

Example (2)

- The posterior distribution up to proportionality is,

$$\pi(\mu,\sigma^2|\boldsymbol{x}) \propto f(\boldsymbol{x}|\mu,\sigma^2)p(\mu)p(\sigma^2)$$

$$\propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right)$$

$$\times \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

$$\propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\times \exp\left(-\frac{\beta}{\sigma^2}\right) \exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right).$$

Example (3)

- Now, the posterior full conditional distributions for μ and σ^2 will be proportional to,

$$\pi(\mu|\sigma, \boldsymbol{x}) \propto exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right)$$

$$\pi(\sigma^2|\mu, \boldsymbol{x}) \propto (\sigma^2)^{-(n/2+\alpha+1)}\exp\left(-\frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}+\beta\right)\right).$$

Example (4)

- The full conditional of μ is a Normal distribution, as we have already shown in a previous lecture that the posterior for μ when σ is known is Normal. You could show this again by completing the square and doing exactly the same algebra we did for Lecture 9.
- The full conditional for σ^2 can easily be recognised as an Inverse Gamma density.
- Therefore, the posterior full conditional distributions for μ and σ^2 are of standard form, namely,

$$\begin{aligned} \mu|\boldsymbol{x}, \sigma^2 &\sim N\left(\frac{\tau^2 n\bar{x} + \sigma^2\phi}{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n + \sigma^2}\right); \\ \sigma^2|\boldsymbol{x}, \mu &\sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta\right), \end{aligned}$$

Example (5)

- Finally, check the R code uploaded on Moodle that implements this Gibbs sampler.