

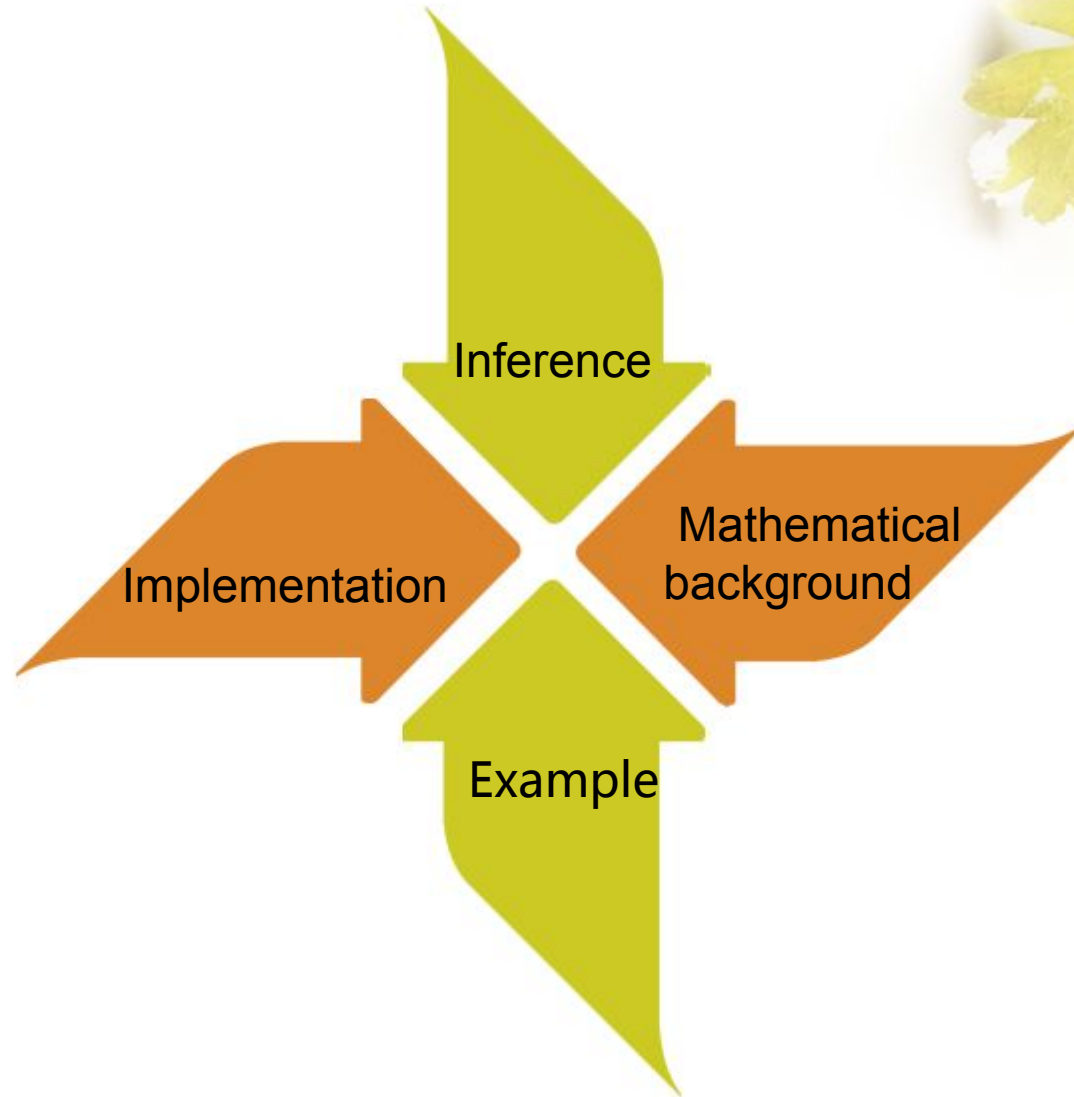
Plum Blossom  
© 2008

# Gibbs Sampling

Made by : Li yongchun  
Yu longhui



# Basic



Improve

Efficient Gibbs sampler

Example

Failure modes

# Basic

## Inference:

Gibbs sampling is commonly used for statistical inference

e.g.

1. determining the best value of a parameter.
2. determining the number of people likely to shop at a particular store on a given day.
3. the candidate a voter will most likely vote for.

etc.

The idea is that observed data is incorporated into the sampling process by creating separate variables for each piece of observed data and fixing the variables in question to their observed values, rather than sampling from those variables.

## Implementation:

Suppose we want to obtain  $k$  samples of  $X = (x_1, x_2, \dots, x_n)$  from a joint distribution  $P(x_1, x_2, \dots, x_n)$ . Denote the  $i$ th sample by  $X^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$ . We proceed as follows:


1. We begin with some initial value  $X^{(0)}$

2. For each sample  $i \in \{1 \dots k\}$ , sample each variable  $x_j^{(i)}$  from the conditional distribution  $p(x_j^{(i)} | x_1^{(i)}, \dots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \dots, x_n^{(i-1)})$

That is, sample each variable from the distribution of that variable conditioned on all other variables, making use of the most recent values and updating the variable with its new value as soon as it has been sampled.

Relation of conditional distribution and joint distribution:


$$p(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n) = \frac{p(x_1, \dots, x_n)}{p(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)} \propto p(x_1, \dots, x_n)$$



The expected value of any variable can be approximated by averaging over all the samples.

1. The initial values of the variables can be determined randomly or by some other algorithm such as expectation-maximization.
2. It is not actually necessary to determine an initial value for the first variable sampled.
3. It is common to ignore some number of samples at the beginning (the so-called *burn-in period*), and then consider only every  $t$ th sample when averaging values to compute an expectation.

For example, the first 1,000 samples might be ignored, and then every 100th sample averaged, throwing away all the rest.



## Mathematical background:

The way to generate a Gibbs sequence of random variables, such as is to start from an initial value  $X'_0 = x'_0, Y'_0 = y'_0$  and iteratively obtain

$X'_0, Y'_0, X'_1, Y'_1 \dots X'_k, Y'_k$  by alternately generating values from

$$X'_j \sim f_{X|Y}(x | Y'_j = y'_j)$$

$$Y'_{j+1} \sim f_{Y|X}(y | X'_j = x'_j)$$

then We briefly explain how and why the Gibbs sampler works.

$f_X(x)$  can be determined as follows:

$$f_X(x) = \int f(x, y) dy$$

- then

$$f_X(x) = \int f_Y(y) f_{X|Y}(x|y) dy$$

- Using similar argument for  $f_Y(y)$ , then

$$f_X(x) = \int \left[ \int f_{X|Y}(x|y) f_{Y|X}(y|t) dy \right] f_X(t) dt$$

- Defines a fixed point integral equation for which  $f_X(x)$  is the solution and the solution is unique.



- Example:

- Table 11.2 Coagulation time in seconds for blood drawn from 24 animals randomly allocated to four different diets.

Diet	Measurements
A	62, 60, 63, 59
B	63, 67, 71, 64, 65, 66
C	68, 66, 71, 67, 68, 68
D	56, 62, 60, 61, 63, 64, 63, 59

- HIERARCHICAL NORMAL MODEL:

- 1. Conditional posterior distribution of each  $\theta_j$ .

$$\theta_j | \mu, \sigma, \tau, y \sim N(\hat{\theta}_j, V_{\theta_j}),$$

$$\hat{\theta}_j = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}} \quad V_{\theta_j} = \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

- 2. Conditional posterior distribution of  $\mu$  :

$$\mu|\theta, \sigma, \tau, y \sim N(\hat{\mu}, \tau^2/J), \quad \hat{\mu} = \frac{1}{J} \sum_{j=1}^J \theta_j.$$

- 3. Conditional posterior distribution of  $\sigma^2$  :

$$\sigma^2|\theta, \mu, \tau, y \sim \text{Inv-}\chi^2(n, \hat{\sigma}^2), \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2.$$

- 4. Conditional posterior distribution of  $\tau^2$

$$\tau^2|\theta, \mu, \sigma, y \sim \text{Inv-}\chi^2(J-1, \hat{\tau}^2), \quad \hat{\tau}^2 = \frac{1}{J-1} \sum_{j=1}^J (\theta_j - \mu)^2.$$

we can choose starting points for each parameter :

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} y_{ij} \quad \hat{\tau}^2 = \hat{\sigma}^2 = \sqrt{\frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2}{n-1}}$$

- we can do 1 step to 4 step for 1000 times, and we can get result like that:

Estimand	Posterior quantiles					$\hat{R}$
	2.5%	25%	median	75%	97.5%	
$\theta_1$	58.9	60.6	61.3	62.1	63.5	1.01
$\theta_2$	63.9	65.3	65.9	66.6	67.7	1.01
$\theta_3$	66.0	67.1	67.8	68.5	69.5	1.01
$\theta_4$	59.5	60.6	61.1	61.7	62.8	1.01
$\mu$	56.9	62.2	63.9	65.5	73.4	1.04
$\sigma$	1.8	2.2	2.4	2.6	3.3	1.00
$\tau$	2.1	3.6	4.9	7.6	26.6	1.05

$$B = \frac{n}{m-1} \sum_{j=1}^m (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2, \text{ where } \bar{\psi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n \psi_{ij}, \quad \bar{\psi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^m \bar{\psi}_{\cdot j}$$

$$W = \frac{1}{m} \sum_{j=1}^m s_j^2, \text{ where } s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (\psi_{ij} - \bar{\psi}_{\cdot j})^2. \quad \widehat{\text{var}}^+(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B.$$

- $\hat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\psi|y)}{W}}$  which declines to 1 as  $n \rightarrow \infty$ .

## Improve

- Efficient Gibbs sampler :

The Gibbs sampler is most efficient when parameterized in terms of independent components. so, we need to reduce the autocorrelation between samples.

### 1. Auxiliary variables

Gibbs sampler computations can often be simplified or convergence accelerated by adding auxiliary variables.

For example:

Fitting the t model (continued), convergence will be slow if a simulation draw of  $\sigma$  is close to zero.

$$\begin{array}{l} y_i \sim N(\mu, V_i) \\ V_i \sim \text{Inv-}\chi^2(\nu, \sigma^2), \end{array} \quad \rightarrow \quad \begin{array}{l} y_i \sim N(\mu, \alpha^2 U_i) \\ U_i \sim \text{Inv-}\chi^2(\nu, \tau^2), \end{array} \quad \alpha > 0$$

$\alpha$  is the auxiliary variable.

## 2. Blocked Gibbs sampler:

A blocked Gibbs sampler groups two or more variables together and samples from their joint distribution conditioned on all other variables, rather than sampling from each one individually.

For example:

In a hidden Markov model, a blocked Gibbs sampler might sample from all the latent variables making up the Markov chain in one go, using the forward-backward algorithm.

remark: A hidden Markov model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states.

### 3. Collapsed Gibbs sampler

A collapsed Gibbs sampler integrates out (marginalizes over) one or more variables when sampling for some other variable.

For example:

imagine that a model consists of three variables A, B, and C. A simple Gibbs sampler

$$\begin{array}{l} p(A, B, C) \\ p(B, A, C) \\ p(C, A, B) \end{array} \rightarrow \begin{array}{l} p(A|C) \\ p(C|A) \end{array} \quad \text{with variable B integrated out in this case}$$

variable B could be collapsed out entirely.

- Example:

Modeling the t distribution as a mixture of normals.

The t likelihood for each data point is equivalent to the model:

$$\begin{aligned}y_i &\sim N(\mu, V_i) \\V_i &\sim \text{Inv-}\chi^2(\nu, \sigma^2),\end{aligned}$$

1. Conditional posterior distribution of each  $V_i$ .

$$V_i | \mu, \sigma^2, \nu, y \sim \text{Inv-}\chi^2\left(\nu + 1, \frac{\nu\sigma^2 + (y_i - \mu)^2}{\nu + 1}\right).$$

2. Conditional posterior distribution of  $\mu$ .

$$\mu | \sigma^2, V, \nu, y \sim N\left(\frac{\sum_{i=1}^n \frac{1}{V_i} y_i}{\sum_{i=1}^n \frac{1}{V_i}}, \frac{1}{\sum_{i=1}^n \frac{1}{V_i}}\right).$$

3. Conditional posterior distribution of  $\sigma^2$ .

$$\sigma^2 | \mu, V, \nu, y \sim \text{Gamma}\left(\frac{n\nu}{2}, \frac{\nu}{2} \sum_{i=1}^n \frac{1}{V_i}\right)$$

- Parameter expansion:
- convergence will be slow if a simulation draw of  $\sigma$  is close to zero

$$y_i \sim N(\mu, \alpha^2 U_i)$$

$$U_i \sim \text{Inv-}\chi^2(\nu, \tau^2),$$

- 1. For each  $i$ ,  $U_i$  is updated much as  $V_i$  was before:

$$U_i | \alpha, \mu, \tau^2, \nu, y \sim \text{Inv-}\chi^2 \left( \nu + 1, \frac{\nu \tau^2 + ((y_i - \mu)/\alpha)^2}{\nu + 1} \right).$$

- 2. The mean,  $\mu$ , is updated as before:

$$\mu | \alpha, \tau^2, U, \nu, y \sim N \left( \frac{\sum_{i=1}^n \frac{1}{\alpha^2 U_i} y_i}{\sum_{i=1}^n \frac{1}{\alpha^2 U_i}}, \frac{1}{\sum_{i=1}^n \frac{1}{\alpha^2 U_i}} \right).$$

- 3. The variance parameter  $\tau^2$ , is updated much as  $\sigma^2$  was before:

$$\tau^2 | \alpha, \mu, U, \nu, y \sim \text{Gamma} \left( \frac{n\nu}{2}, \frac{\nu}{2} \sum_{i=1}^n \frac{1}{U_i} \right)$$

- 4. Finally, we must update  $\alpha^2$ , which is easy since conditional on all the other parameters in the model it is simply a normal variance parameter:

$$\alpha^2 | \mu, \tau^2, U, \nu, y \sim \text{Inv-}\chi^2 \left( n, \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \mu)^2}{U_i} \right).$$



- Failure modes:

There are two ways that Gibbs sampling can fail.

1. The first is when there are islands of high-probability states, with no paths between them. Gibbs sampling will become trapped in one of the two high-probability vectors, and will never reach the other one.

For example:

Consider a probability distribution over 2-bit vectors.

(0,0)

(1,0)

(0,1)

(1,1)

$P = 1/2$

$P = 0$

$P = 0$

$P = 1/2$

Gibbs sampling will become trapped in one of the two high-probability vectors, and will never reach the other one.

More generally, for any distribution over high-dimensional, real-valued vectors, if two particular elements of the vector are perfectly correlated (or perfectly anti-correlated), those two elements will become stuck, and Gibbs sampling will never be able to change them.

1.  $X=(x_1, x_2)$  from  $P(x_1, x_2)$

(0,0)  
 $P=1/2$

(1,0)  
 $P=0$

(0,1)  
 $P=0$

(1,1)  
 $P=1/2$

- 2. (i)  $X^{(0)} = (x_1^{(0)}, x_2^{(0)}) = (0,1)$  or (ii)  $X^{(0)} = (x_1^{(0)}, x_2^{(0)}) = (1,0)$

$$x_1^{(1)} \sim p(x_1^{(1)} | x_2^{(0)}) = p(x_1^{(1)} | 1) = 1$$

$$x_1^{(1)} \sim p(x_1^{(1)} | x_2^{(0)}) = p(x_1^{(1)} | 0) = 0$$

- $x_2^{(1)} \sim p(x_2^{(1)} | x_1^{(1)}) = p(x_2^{(1)} | 1) = 1$

$$x_2^{(1)} \sim p(x_2^{(1)} | x_1^{(1)}) = p(x_2^{(1)} | 0) = 0$$

...

...

$$x_1^{(n)} \sim p(x_1^{(n)} | x_2^{(n-1)}) = p(x_1^{(n)} | 1) = 1$$

$$x_1^{(n)} \sim p(x_1^{(n)} | x_2^{(n-1)}) = p(x_1^{(n)} | 0) = 0$$

$$x_2^{(n)} \sim p(x_2^{(n)} | x_1^{(n)}) = p(x_2^{(n)} | 1) = 1$$

$$x_2^{(n)} \sim p(x_2^{(n)} | x_1^{(n)}) = p(x_2^{(n)} | 0) = 0$$

- 3. we got Gibbs sample,  $\begin{cases} X^{(1)} = X^{(2)} = \dots = X^{(n)} = (1,1) \\ X^{(1)} = X^{(2)} = \dots = X^{(n)} = (0,0) \end{cases}$  OR

- 4. Two particular elements of the vector are perfectly correlated,

$x_1 = x_2$  and Gibbs sampling will never be able to change them.

2. The second problem can happen even when all states have nonzero probability and there is only a single island of high-probability states.


For example:

Consider a probability distribution over 100-bit vectors.

$$P = \begin{cases} \textit{other} & p = \frac{1}{2(2^{100} - 1)}; \\ (0, 0 \dots, 0) & p = 1/2 \end{cases}$$

If you want to estimate the probability of the zero vector, it would be sufficient to take 100 or 1000 samples from the true distribution. That would very likely give an answer very close to  $\frac{1}{2}$ .


But you would probably have to take more than samples from Gibbs sampling to get the same result. No computer could do this in a lifetime.



Gibbs sampling will alternate between returning only the zero vector for long periods (about  $10^6$  in a row), then only nonzero vectors for long periods (about  $10^6$  in a row).

Thus convergence to the true distribution is extremely slow, requiring much more than  $10^6$  steps; taking this many steps is not computationally feasible in a reasonable time period.

Note that a problem like this can be solved by block sampling the entire 100-bit vector at once.



---

站长素材  
sc.chinaz.com

Thank you !

