Bayesian Mixture Models and the Gibbs Sample

Final report of Appiled Bayesian Methods

109354015 楊承鑫
110354017 趙立騰

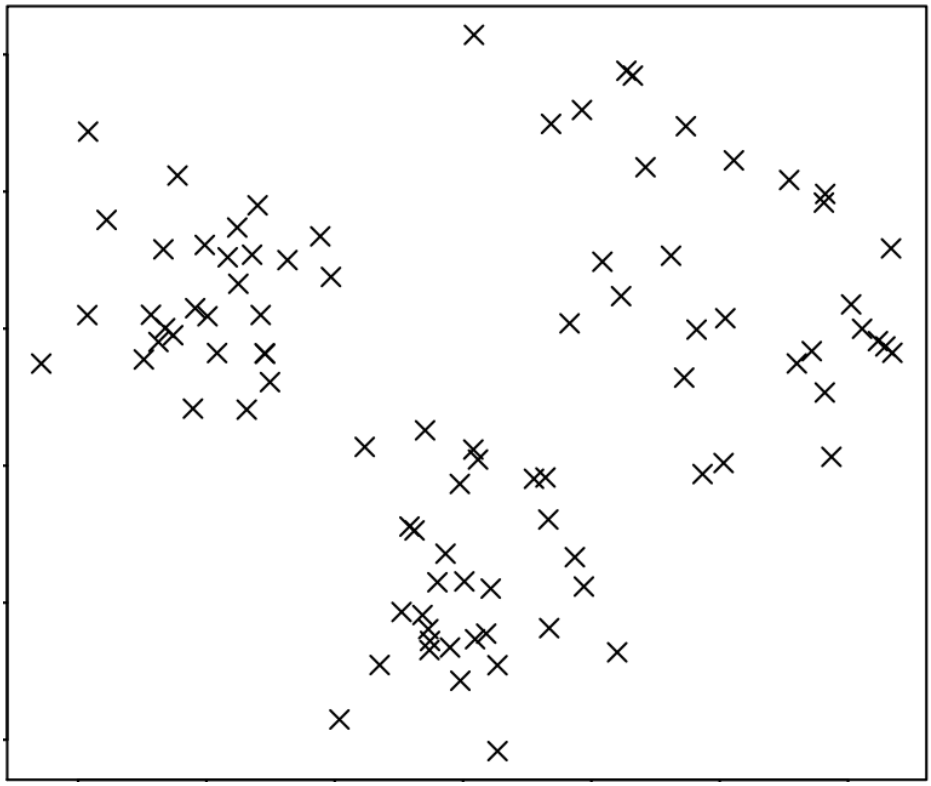
Bayesian mixture of Gaussians

$$X_i|Z_i \sim N\left(\mu_{Z_i}, \Sigma_{Z_i}\right) \text{ , } i = 1, \dots, n, \text{independently}$$

$$p(Z_i = k) = \pi_k \text{ , } k = 1, \dots, K \text{ , } i = 1, \dots, n, \text{independently}$$

X_i : data point (observed), $i = 1, \dots, n$
Z_i : cluster assignment(unobserved) , $i = 1, \dots, n$
μ_k : cluster location , $k = 1, \dots, K$
Σ_k : cluster variance-covariance matrix , $k = 1, \dots, K$

This is a plot of data points:



Plot 1.

Posterior predictive distribution

$$\mu_k \sim N(\mu_{k0}, \Sigma_{k0}) \text{ , k} = 1, \dots, \text{K,independently}$$

$$p(x_{n+1}|\vec{x}) = \int_{\vec{\mu}} p(x_{n+1}|\vec{\mu}) p(\vec{\mu}|\vec{x})\, d\vec{\mu}$$

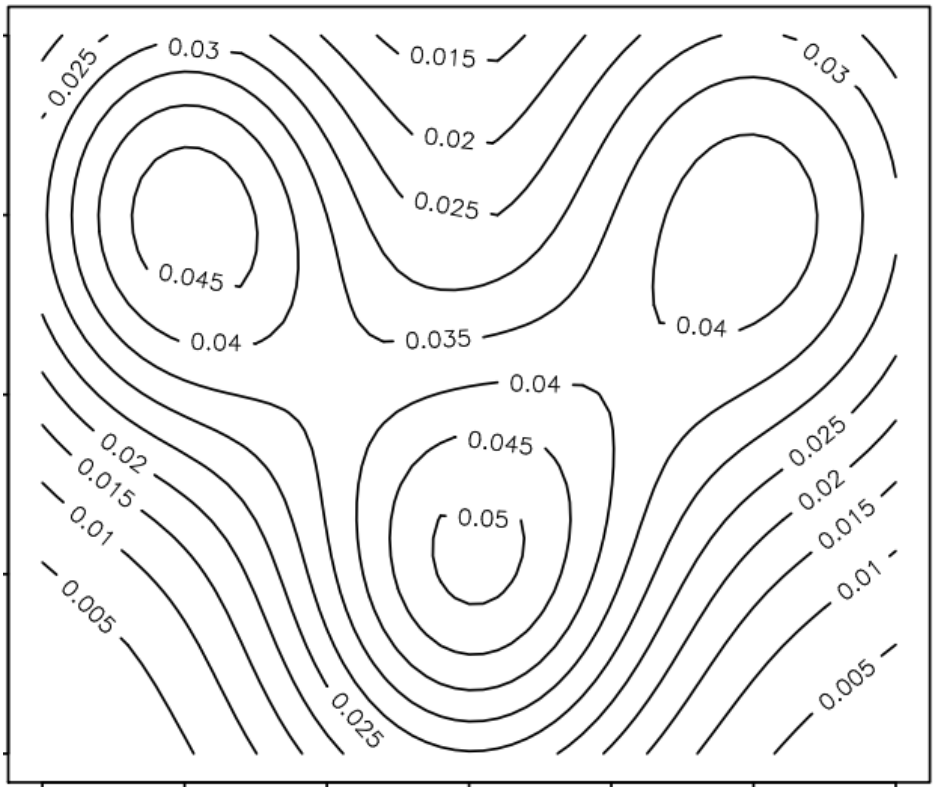
$$= \int_{\vec{\mu}} \left(\sum_{k=1}^{K} p(Z_{n+1} = k) p(x_{n+1}|\mu_k) \right) p(\vec{\mu}|\vec{x})\, d\vec{\mu}$$

$$= \sum_{k=1}^{K} p(Z_{n+1} = k) \int_{\vec{\mu}} p(x_{n+1}|\mu_k) p(\vec{\mu}|\vec{x})\, d\vec{\mu}$$

$$= \sum_{k=1}^{K} p(Z_{n+1} = k) \int_{\mu_k} p(x_{n+1}|\mu_k) p(\mu_k|\vec{x})\, d\mu_k \qquad (1)$$

Therefore,we can consider x_{n+1} as coming from each of the potential cluster locations and then take a weighted average with the weight being equal to the corresponding probabilities.

Here is a schematic plot of the predictive distribution based on the data points in plot 1:



Plot 2.

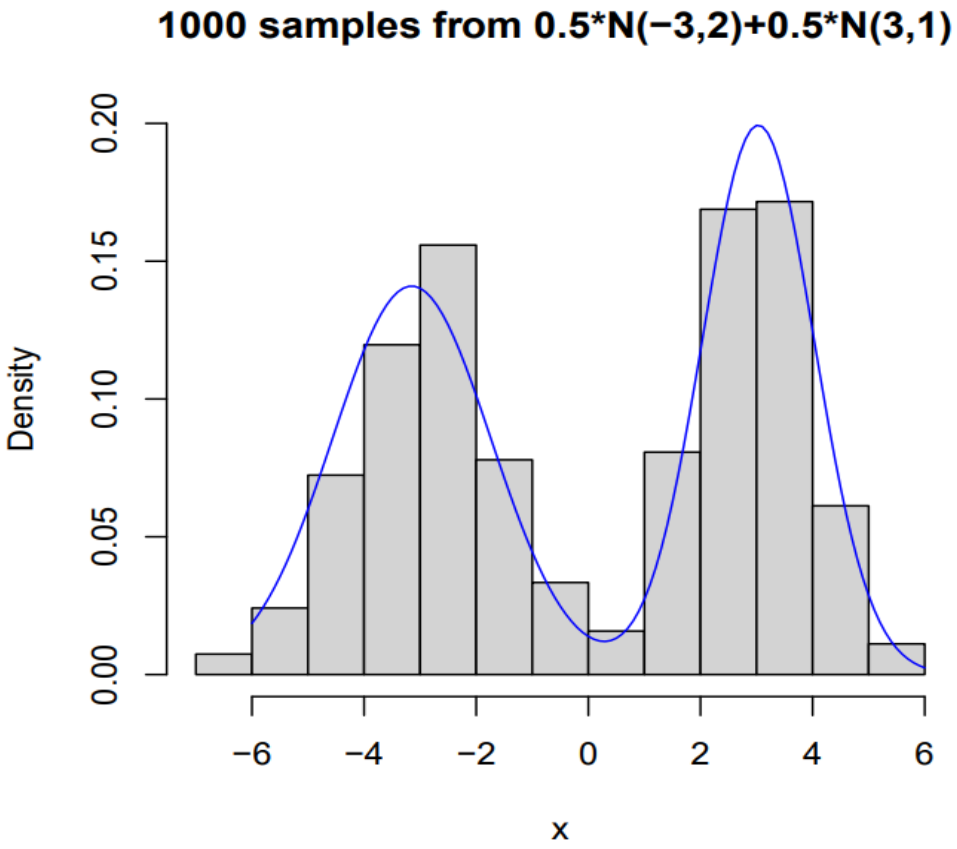
Next we focus on the 1-dim model,

$$X_i|Z_i \sim N\left(\mu_{Z_i}, {\sigma^2}_{Z_i}\right) \ , i = 1, \dots, n, \text{independently}$$

$$p(Z_i = k) = \pi_k \ , k = 1, \dots, K, i = 1, \dots, n, \text{independently}$$

$$\mu_k \sim N(\mu_{k0}, \lambda^2) \ , \mathrm{k} = 1, \dots, K, \text{independently}$$

Example



The blue line is a posterior predictive distribution based on the 1000 samples.To compute predictive distribution (1) ,we need that data is given. Howerer,when computing $p(x|\vec{\mu}) = \sum_{k=1}^{K} p(Z = k)p(x|\mu_k)$,we require that $\vec{\mu}$ is given.And this is a considerable difference between the collapsed Gibbs sampling and the Gibbs sampling we mention below.

The Gibbs sampler

- The Gibbs sampler is Metropolis-Hastings with taking full conditional as

proposal distribution.Then the acceptance probability equals 1.

- The samples from a Markov chain take the target distribution as its stationary distribution.

- Using a set of samples to approximate a distribution:

$$p(\mu, z|x) \approx \frac{1}{B}\sum_{b=1}^{B} \delta_{(\mu^{(b)},z^{(b)})}(\mu, z)$$

- If full conditionals are easier to compute,it is worthwhile to use Gibbs sampling

The full conditionals are

$p(Z_i = k|\vec{\mu}, \vec{x}) = \frac{\pi_k \phi(x_i;\mu_k,\sigma_k^2)}{\sum_{k=1}^{K} \pi_k \phi(x_i;\mu_k,\sigma_k^2)}$,where $\phi(\bullet;\mu,\sigma^2)$ is a normal pdf with mean μ and variance $\sigma^2, \mathrm{k} = 1,..,\mathrm{K}\,, \mathrm{i} = 1,\ldots,\mathrm{n}$. (2)

(see more details in appendix)

We don't need given z_{-i} because of the conditional independence between any z_i and z_j for a given $\vec{\mu}$.

$$\mu_k|\vec{z},\vec{x} \sim N\left(\frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\mu_{0k}\,,\ \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\right), \text{where } n_k = \#\,of\,\{i : z_i = k\} \text{ and } \bar{x}_k = \Sigma_{i:z_i=k}\frac{x_i}{n_k}, \mathrm{k} = 1,..,\mathrm{K} \quad (3)$$

(see more details in appendix)

Algorithm

1. **Given the initial value $\vec{\mu}$**
2. **For each i∈{1,…,n} : Sample z_i from equation (2).**
3. **For each k∈{1,…,K} : Sample μ_k from equation (3).**

And then repeat 2.&3.

The collapsed Gibbs sampler

- Integrating out hidden random variables from a full conditional is called ***collapsing***.

- It can be seem as a predictive updated version of the Gibbs sampler.

- Althought it costs more than Gibbs sampling per iteration,it othen result in faster convergence of the Markov chain to the stationary distribution.

In our model,we can collapse the mixture locations,

$$p(z_i = k|z_{-i}, \vec{x}) \propto p(z_i = k)p(x_i|z_{-i}, x_{-i}, z_i = k) \qquad (4)$$

,where

$$p(x_i|z_{-i}, x_{-i}, z_i = k) = \int_{\mu_k} p(x_i|\mu_k)\ p(\mu_k|z_{-i}, x_{-i}) \qquad (5)$$

Equation (5) is simply a posterior predictive distribution.

Based on (3),(4),and(5),we can obtain

$$p(z_i = k|z_{-i}, \vec{x})$$

$$= \frac{\pi_k \phi\left(x_i; \frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\mu_{0k}, \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} + \sigma_k^2\right)}{\sum_{k=1}^{K} \pi_k \phi\left(x_i; \frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\mu_{0k}, \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} + \sigma_k^2\right)}$$

where $n_k = \#\ of\ \{a: z_a = k\}$ and $\bar{x}_k = \Sigma_{a:z_a=k} \frac{x_a}{n_k}$ with a belonging to $\{1, \dots, n\} - \{i\}, \mathrm{k} = 1, \dots, \mathrm{K}, \mathrm{i} = 1, \dots, \mathrm{n}$ (6)

(see more details in appendix)

Algorithm

1. **Given the initial value $\vec{z}$**
2. **Sample z_1 from equation (6).**

3. Sample z_2 from equation (6).

.

.

and so on. Until sampling z_n from equation (6),we finish one iteration. Take the new $\vec{z}$ as the initial value and repeat the procedure from 1.

Comparison

We use **eruptions data** recording that the duration of each of eruptions for the Old Faithful geyser in Yellowstone National Park from August 1 to August 15, 1985 to compare the Gibbs sampler with the collapsed Gibbs sampler.

The first eight cluster assignments(iteration=20)

The Gibbs sampler

Inference for the input samples (3 chains: each with iter = 20; warmup = 10):

	Q5	Q50	Q95	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
V1	2.0	2	2.0	2.0	0.2	1.00	15	30
V2	1.0	1	1.0	1.0	0.0	1.00	30	30
V3	1.0	2	2.0	1.8	0.4	1.06	15	30
V4	1.0	1	1.5	1.1	0.3	0.98	15	15
V5	2.0	2	2.0	2.0	0.0	1.00	30	30
V6	1.0	1	2.0	1.4	0.5	1.06	15	30
V7	2.0	2	2.0	2.0	0.0	1.00	30	30
V8	1.5	2	2.0	1.9	0.3	0.98	15	30

Table1.

The collapsed Gibbs sampler

Inference for the input samples (3 chains: each with iter = 20; warmup = 10):

	Q5	Q50	Q95	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
V1	2	2	2	2.0	0.0	1.00	30	30
V2	1	1	1	1.0	0.0	1.00	30	30
V3	1	2	2	1.8	0.4	0.98	15	30
V4	1	1	1	1.0	0.0	1.00	30	30
V5	2	2	2	2.0	0.0	1.00	30	30
V6	1	1	2	1.1	0.3	1.01	15	30
V7	2	2	2	2.0	0.0	1.00	30	30
V8	2	2	2	2.0	0.0	1.00	30	30

Table2.

The first eight cluster assignments(iteration=2000)

The Gibbs sampler

Inference for the input samples (3 chains: each with iter = 2000; warmup = 1000):

	Q5	Q50	Q95	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
V1	2	2	2	2.0	0.2	1	2988	3000
V2	1	1	1	1.0	0.0	1	3018	3018
V3	1	2	2	1.9	0.3	1	2936	3000
V4	1	1	1	1.0	0.2	1	3068	3068
V5	2	2	2	2.0	0.0	1	3000	3000
V6	1	1	2	1.4	0.5	1	2899	3000
V7	2	2	2	2.0	0.0	1	3000	3000
V8	2	2	2	2.0	0.2	1	2912	3000

Table3.

The collapsed Gibbs sampler

Inference for the input samples (3 chains: each with iter = 2000; warmup = 1000):

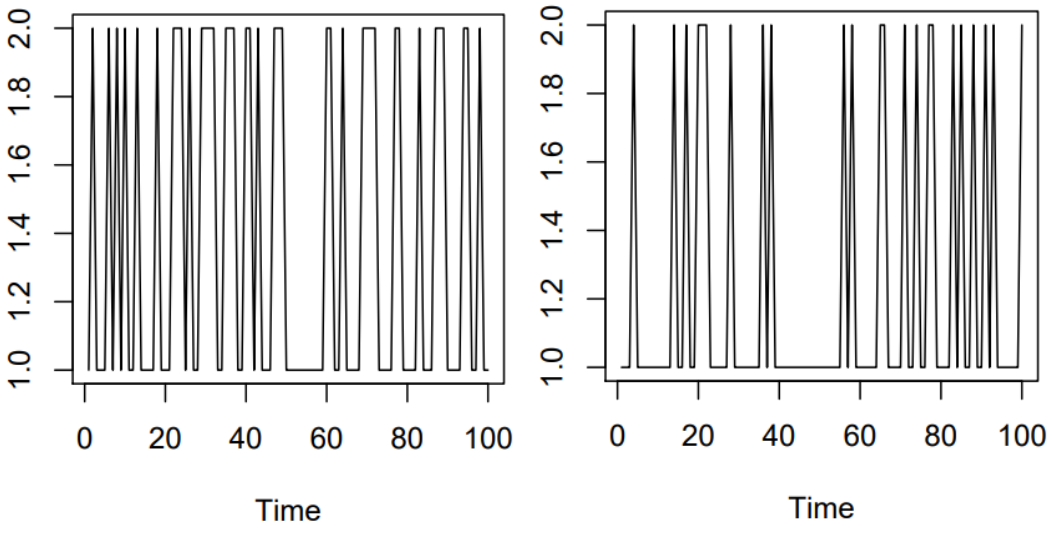
	Q5	Q50	Q95	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
V1	2	2	2	2.0	0.1	1	2883	3000
V2	1	1	1	1.0	0.0	1	3000	3000
V3	1	2	2	1.9	0.2	1	3049	3000
V4	1	1	1	1.0	0.0	1	3026	3026
V5	2	2	2	2.0	0.0	1	3000	3000
V6	1	1	2	1.3	0.5	1	3209	3000
V7	2	2	2	2.0	0.0	1	3000	3000
V8	2	2	2	2.0	0.1	1	3037	3000

Table4.

According to **table1.**,when setting the threshold of Rhat of convergence is 1.05 and number of iterations is 20 ,we can observe that the Gibbs sampling result in divergence of z_3 and z_6**.** However**, table2.**(the collapsed Gibbs sampler) indicates that all of the first eight cluster assignments are convergent.Hence,the collapsed Gibbs sampling seems to result in faster convergence than the Gibbs sampling.
From **table3.,**and **table4.**,there is no considerable difference between the two methods when number of iterations is 2000.The standard errors(SD) of the collapsed Gibbs sampling are generally smaller than those of the Gibbs sampling.

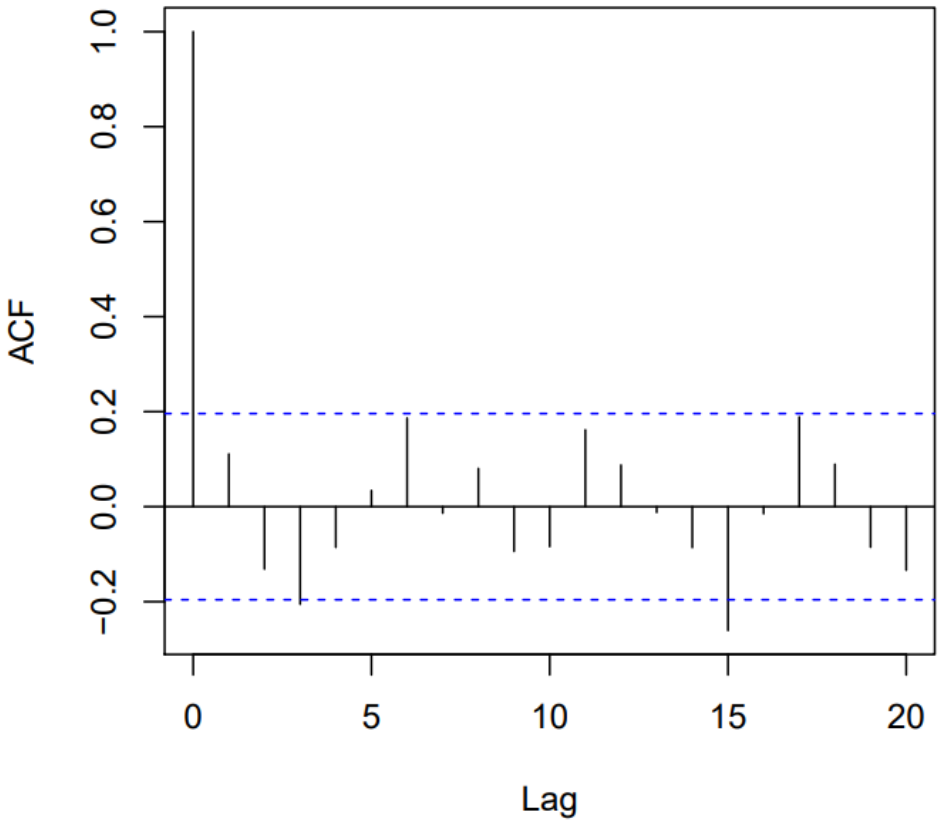
Time series and ACF of z_6 (iteration=100)

The Gibbs sampler **The collapsed Gibbs sampler**



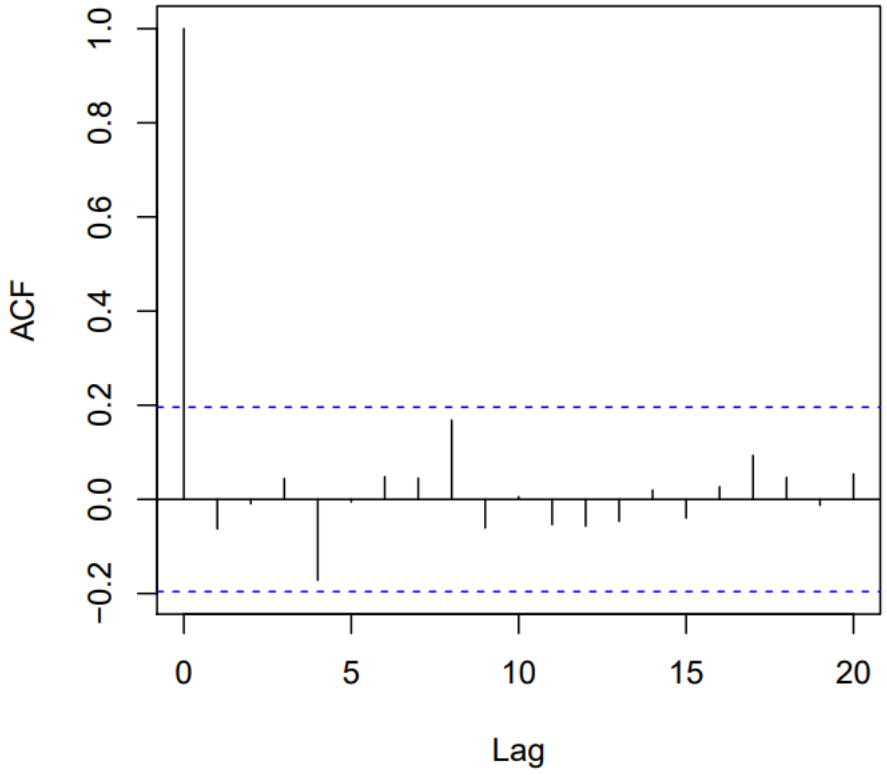
Plot3. **Plot4.**

The Gibbs sampler



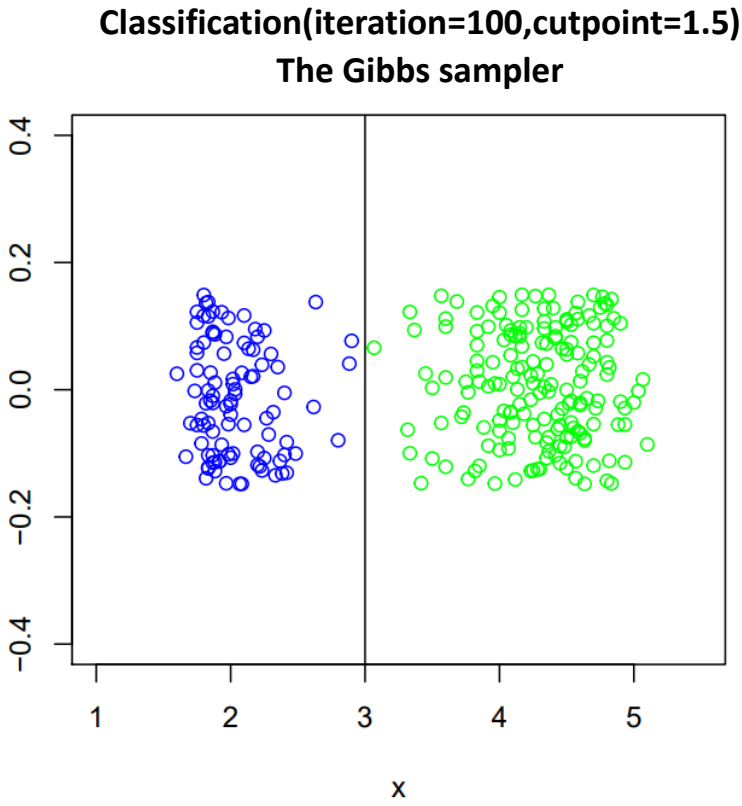
Plot5.

The collapsed Gibbs sampler

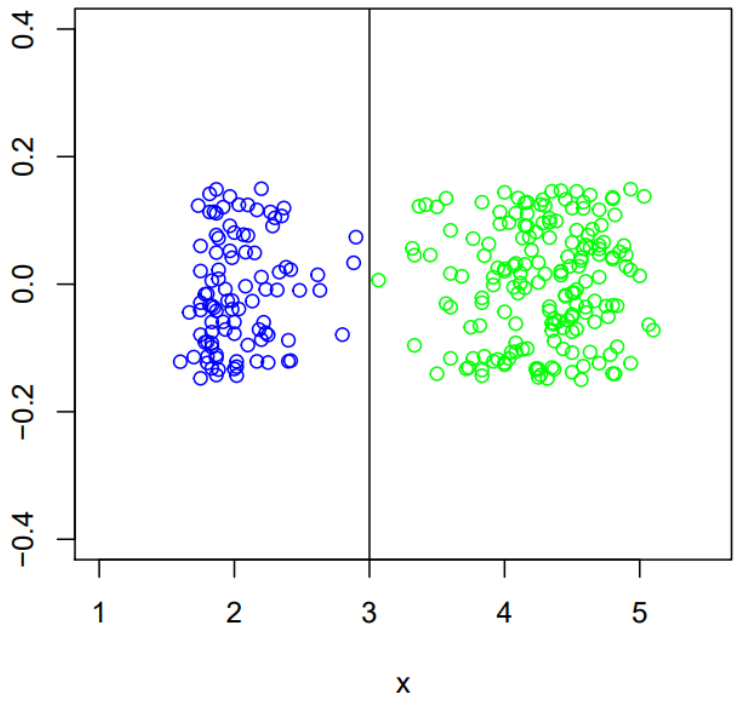


Plot6.

From **Plot5.** and **Plot6.**,by contrast to the Gibbs sampling,the collapsed Gibbs sampling reduces some autocorrelations.



Plot7.



Plot8.

The dots in **Plot7.** and **Plot8.** are shifted randomly along y axis(jittered). From **Plot7.** and **Plot8.**,the result of classification for the collapsed Gibbs sampler is the same as that for the Gibbs sampler.

Appendix

(a) **Proof of** (2):

$$p(z_i|\vec{\mu},\vec{x}) = p(z_i|z_{-i},\vec{\mu},\vec{x}) \propto \prod_{\mathbf{i=1}}^{\mathbf{n}} p(z_i,x_i|\vec{\mu}) \prod_{\mathbf{k=1}}^{\mathbf{K}} \phi(\mu_k;\mu_{k0},\lambda^2)$$

$$\propto \prod_{\mathbf{i=1}}^{\mathbf{n}} p(z_i,x_i|\vec{\mu}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \pi_{z_i}\,\phi\left(x_i;\mu_{z_i},{\sigma^2}_{z_i}\right)$$

$$\propto \pi_{z_i}\,\phi\left(x_i;\mu_{z_i},{\sigma^2}_{z_i}\right)\ ,i=1,\dots,n$$

$p(Z_i = k|\vec{\mu},\vec{x}) = C\pi_k\,\phi(x_i;\mu_k,{\sigma^2}_k), k=1,\dots,K, i=1,\dots,n$, where C is a normalized constant.

$$\Rightarrow C = \frac{1}{\sum_{k=1}^{K}\pi_k\,\phi(x_i;\mu_k,{\sigma^2}_k)}, \mathrm{i}=1,\dots,n$$

$$\Rightarrow\ p(Z_i = k|\vec{\mu},\vec{x}) = \frac{\pi_k\phi(x_i;\mu_k,\sigma_k^2)}{\sum_{k=1}^{K}\pi_k\phi(x_i;\mu_k,\sigma_k^2)}\ , k=1,\dots,K\ , \mathrm{i}=1,\dots,n$$

(b) **Proof of** (3):

$$p(\mu_k|\vec{z},\vec{x}) = p(\mu_k|\mu_{-k},\vec{z},\vec{x}) \propto \prod_{\mathbf{i=1}}^{\mathbf{n}} p(z_i,x_i|\vec{\mu}) \prod_{\mathbf{k=1}}^{\mathbf{K}} \phi(\mu_k;\mu_{k0},\lambda^2)$$

$$\propto \prod_{\mathbf{i=1}}^{\mathbf{n}} \pi_{z_i}\,\phi\left(x_i;\mu_{z_i},{\sigma^2}_{Z_i}\right) \prod_{\mathbf{k=1}}^{\mathbf{K}} \phi\left(\mu_k;\mu_{k0},\lambda^2\right)$$

$$\propto \left(\prod_{i:z_i=k}\phi\left(x_i;\mu_k,{\sigma^2}_k\right)\right)\phi\left(\mu_k;\mu_{k0},\lambda^2\right), \mathrm{k}=1,..,K$$

Hence, $\mu_k|\vec{z},\vec{x} \sim N\left(\frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\mu_{0k}\ ,\ \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\right), \mathrm{k}=1,..,K$,where

$n_k = \#\ of\ \{i : z_i = k\}$ and $\bar{x}_k = \Sigma_{i:z_i=k} \frac{x_i}{n_k}$.

(c) **Proof of** (6):
Based on (4)&(5),we can obtain

$$p(z_i = k|z_{-i}, \vec{x}) \propto p(z_i = k) \int_{\mu_k} p(x_i|\mu_k)\ p(\mu_k|z_{-i}, x_{-i})$$

$$\propto p(z_i = k) \int_{\mu_k} \phi\left(x_i; \mu_k, {\sigma^2}_k\right)\ \phi\left(\mu_k; \frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \mu_{0k}, \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}}\right)$$ (according to (3))

$$\propto p(z_i = k)\ \phi\left(x_i; \frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \mu_{0k}, \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} + \sigma^2{}_k\right)\ k = 1, \dots, K, i = 1, \dots, n,$$

where $n_k = \#\ of\ \{a: z_a = k\}$ and $\bar{x}_k = \Sigma_{a:z_a=k} \frac{x_a}{n_k}$ with a belonging to $\{1, \dots, n\} - \{i\}$.

Hence,

$$p(z_i = k|z_{-i}, \vec{x})$$
$$= \frac{\pi_k \phi\left(x_i; \frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \mu_{0k}, \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} + \sigma_k^2\right)}{\sum_{k=1}^{K} \pi_k \phi\left(x_i; \frac{\frac{n_k}{\sigma_k^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \bar{x}_k + \frac{\frac{1}{\lambda^2}}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} \mu_{0k}, \frac{1}{\frac{n_k}{\sigma_k^2}+\frac{1}{\lambda^2}} + \sigma_k^2\right)}$$

, $\mathrm{k} = 1, \dots, K$, $\mathrm{i} = 1, \dots, n$,where $n_k = \#\ of\ \{a: z_a = k\}$ and $\bar{x}_k = \Sigma_{a:z_a=k} \frac{x_a}{n_k}$ with a belonging to $\{1, \dots, n\} - \{i\}$

Reference

1. Jun S. Liu(1994), The Collapsed Gibbs Sampler in Bayesian

Computations with Applications to a Gene Regulation Problem, *Journal of the American Statistical Association , Sep., 1994, Vol. 89, No. 427,* page. 958-966

2. David M. Blei(2015), Bayesian Mixture Models and the Gibbs Sampler,Oct.,2015

3. Course slides for Applied Bayesian Methods