Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca
1.1– Outline

- The Gibbs Sampler
- Variable Selection Example
- Finite Mixture of Gaussians
2.1– Summary of Last Lecture

• The Gibbs sampler is a generic method to sample from high-dimensional distribution.

• It generates a Markov chain which converges to the target distribution under weak assumptions: irreducibility and aperiodicity.
2.2– More about the Gibbs sampler

- If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.

- Initialization:
  - Select deterministically or randomly $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$.

- Iteration $i; i \geq 1$:
  
  For $k = 1 : p$

  - Sample $\theta_k^{(i)} \sim \pi \left( \theta_k | \theta_{-k}^{(i)} \right)$.

  where $\theta_{-k}^{(i)} = (\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)})$. 
2.3– Random Scan Gibbs sampler

- Initialization:
  - Select deterministically or randomly $\theta^{(0)} = \left( \theta_1^{(0)}, \ldots, \theta_p^{(0)} \right)$.

- Iteration $i; \ i \geq 1$:
  - Sample $K \sim U_{\{1, \ldots, p\}}$.
  - Set $\theta^{(i)}_{-K} = \theta^{(i-1)}_{-K}$.
  - Sample $\theta^{(i)}_K \sim \pi \left( \theta_K | \theta^{(i)}_{-K} \right)$.

where $\theta^{(i)}_{-K} = \left( \theta_1^{(i)}, \ldots, \theta_{K-1}^{(i)}, \theta_{K+1}^{(i)}, \ldots, \theta_p^{(i)} \right)$. 

– Summary
2.4– Practical Recommendations

- Try to have as few “blocks” as possible.

- Put the most correlated variables in the same block.

- If necessary, reparametrize the model to achieve this.

- Integrate analytically as many variables as possible: pretty algorithms can be much more inefficient than ugly algorithms.

- There is no general result telling strategy A is better than strategy B in all cases: you need experience.
3.1– Bayesian Variable Selection Example

- We select the following model

\[ Y = \sum_{k=1}^{p} \beta_k X_k + \sigma V \]  

where \( V \sim \mathcal{N}(0,1) \)

where we assume \( IG(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}) \) and for \( \alpha^2 << 1 \)

\[ \beta_k \sim \frac{1}{2} \mathcal{N}(0, \alpha^2 \delta^2 \sigma^2) + \frac{1}{2} \mathcal{N}(0, \delta^2 \sigma^2) \]

- We introduce a latent variable \( \gamma_k \in \{0, 1\} \) such that

\[ \Pr(\gamma_k = 0) = \Pr(\gamma_k = 1) = \frac{1}{2}, \]

\[ \beta_k | \gamma_k = 0 \sim \mathcal{N}(0, \alpha^2 \delta^2 \sigma^2), \quad \beta_k | \gamma_k = 1 \sim \mathcal{N}(0, \delta^2 \sigma^2). \]
• We have parameters \((\beta_{1:p}, \gamma_{1:p}, \sigma^2)\) and observe \(n\) observations \(D = \{x_j, y_j\}_{j=1}^n\).

• A potential Gibbs sampler consists of sampling iteratively from \(p(\beta_{1:p} \mid D, \gamma_{1:p}, \sigma^2)\) (Gaussian), \(p(\sigma^2 \mid D, \gamma_{1:p}, \beta_{1:p})\) (inverse-Gamma) and \(p(\gamma_{1:p} \mid D, \beta_{1:p}, \sigma^2)\).

• In particular

\[
p(\gamma_{1:p} \mid D, \beta_{1:p}, \sigma^2) = \prod_{k=1}^p p(\gamma_k \mid \beta_k, \sigma^2)
\]

and

\[
p(\gamma_k = 1 \mid \beta_k, \sigma^2) = \frac{\frac{1}{\sqrt{2\pi\delta\sigma}} \exp\left(-\frac{\beta_k^2}{2\delta^2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\delta\sigma}} \exp\left(-\frac{\beta_k^2}{2\delta^2\sigma^2}\right) + \frac{1}{\sqrt{2\pi\alpha\delta\sigma}} \exp\left(-\frac{\beta_k^2}{2\alpha^2\delta^2\sigma^2}\right)}.
\]

• The Gibbs sampler becomes reducible as \(\alpha\) goes to zero.
3.3– Bayesian Variable Selection Example

• This is the result of bad modelling and bad algorithm. You would like to put $\alpha \sim 0$ and write

$$Y = \sum_{k=1}^{p} \gamma_k \beta_k X_k + \sigma V \text{ where } V \sim \mathcal{N}(0, 1)$$

where $\gamma_k = 1$ if $X_k$ is included or $\gamma_k = 0$ otherwise. However this suggests that $\beta_k$ is defined even when $\gamma_k = 0$.

• A neater way to write such models is to write

$$Y = \sum_{\{k: \gamma_k = 1\}} \beta_k X_k + \sigma V = \beta_{\gamma}^T X_\gamma + \sigma V$$

where, for a vector $\gamma = (\gamma_1, \ldots, \gamma_p)$, $\beta_\gamma = \{\beta_k : \gamma_k = 1\}$, $X_\gamma = \{X_k : \gamma_k = 1\}$ and $n_\gamma = \sum_{k=1}^{p} \gamma_k$.

• Prior distributions

$$\pi_\gamma (\beta_\gamma, \sigma^2) = \mathcal{N} (\beta_\gamma; 0, \delta^2 \sigma^2 I_{n_\gamma}) \mathcal{IG} \left( \sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2} \right)$$

and $\pi (\gamma) = \prod_{k=1}^{p} \pi (\gamma_k) = 2^{-p}$. 
3.4– A Better Gibbs Sampler

• We are interested in sampling from the trans-dimensional distribution $\pi(\gamma, \beta, \sigma^2 | D)$

• However, we know that

$$\pi(\gamma, \beta, \sigma^2 | D) = \pi(\gamma | D) \pi(\beta, \sigma^2 | D, \gamma)$$

where

$$\pi(\gamma | D) \propto \pi(D | \gamma) \pi(\gamma)$$

and

$$\pi(D | \gamma) = \int \pi(D, \beta, \sigma^2 | \gamma) \, d\beta \, d\sigma^2$$

$$\propto \Gamma\left(\frac{\nu_0 + n}{2} + 1\right) \delta^{-n\gamma} |\Sigma\gamma|^{1/2} \left(\frac{\gamma_0 + \sum_{j=1}^{n} y_{k}^2 - \mu_T \Sigma^{-1} \mu}{2}\right)^{-\left(\frac{\nu_0+n}{2}+1\right)}$$
The full conditional distribution for $\pi \left( \beta_{\gamma}, \sigma^2 \mid D, \gamma \right)$ is

$$
\pi_{\gamma} \left( \beta_{\gamma}, \sigma^2 \mid D \right) = N \left( \beta_{\gamma}; \mu_{\gamma}, \sigma^2 \Sigma_{\gamma} \right)
\times IG \left( \sigma^2; \frac{\nu_0 + n}{2}, \frac{\gamma_0 + \sum_{j=1}^{n} y_j^2 - \mu_{\gamma}^T \Sigma_{\gamma}^{-1} \mu_{\gamma}}{2} \right)
$$

where

$$
\mu_{\gamma} = \Sigma_{\gamma} \left( \sum_{j=1}^{n} y_j x_{\gamma,j} \right), \quad \Sigma_{\gamma}^{-1} = \delta^{-2} I_{n_{\gamma}} + \sum_{i=1}^{n} x_{\gamma,j} x_{\gamma,j}^T.
$$
3.4— A Better Gibbs Sampler

• Popular alternative Bayesian models include

\[ \gamma_i \sim \mathcal{B}(\lambda) \text{ where } \lambda \sim \mathcal{U}[0, 1], \]

\[ \gamma_i \sim \mathcal{B}(\lambda_i) \text{ where } \lambda_i \sim \mathcal{B}e(\alpha, \beta). \]

• g-prior (Zellner)

\[ \beta_\gamma | \sigma^2 \sim \mathcal{N}\left(\beta_\gamma; 0, \delta^2 \sigma^2 \left(X_\gamma^T X_\gamma\right)^{-1}\right). \]

• Robust models where additionally one has

\[ \delta^2 \sim \mathcal{I}G\left(\frac{a_0}{2}, \frac{b_0}{2}\right). \]

• Such variations are very important and can modify dramatically the performance of the Bayesian model.
3.5– Collapsed Gibbs Sampler for Bayesian Variable Selection

- $\pi(\gamma | D)$ is a discrete probability distribution with $2^p$ potential values, we assume $\delta^2$ known here.

- Initialization:
  - Select deterministically or randomly $\gamma^{(0)} = (\gamma_1^{(0)}, ..., \gamma_p^{(0)})$.

- Iteration $i; i \geq 1$:
  - For $k = 1 : p$
    - Sample $\gamma_k^{(i)} \sim \pi(\gamma_k | D, \gamma_{-k}^{(i)})$.
  where $\gamma_{-k}^{(i)} = (\gamma_1^{(i)}, ..., \gamma_{k-1}^{(i)}, \gamma_{k+1}^{(i-1)}, ..., \gamma_p^{(i-1)})$.

  - Optional step: Sample $\left(\beta_{\gamma}^{(i)}, \sigma^2(i)\right) \sim \pi(\beta_{\gamma}, \sigma^2 | D, \gamma^{(i)})$. 
3.6– Bayesian Variable Selection Example

• Consider the case where $\delta^2$ is unknown.

• Initialization:
  • Select deterministically or randomly $\left(\gamma^{(0)}, \beta^{(0)}_\gamma, \sigma^2(0), \delta^2(0)\right)$

• Iteration $i; i \geq 1$:
  For $k = 1 : p$
  • Sample $\gamma_k^{(i)} \sim \pi \left(\gamma_k | D, \gamma_{-k}^{(i)}, \delta^2(i-1)\right)$.
  where $\gamma_{-k}^{(i)} = \left(\gamma_1^{(i)}, \ldots, \gamma_{k-1}^{(i)}, \gamma_{k+1}^{(i-1)}, \ldots, \gamma_p^{(i-1)}\right)$.
  • Sample $\left(\beta^{(i)}_\gamma, \sigma^2(i)\right) \sim \pi \left(\beta_\gamma, \sigma^2 | D, \gamma^{(i)}, \delta^2(i)\right)$.
  • Sample $\delta^2(i) \sim \pi \left(\delta^2(i) | \beta^{(i)}_\gamma\right)$. 
Caterpillar dataset: 1973 study to assess the influence of some forest settlement characteristics on the development of caterpillar colonies.

The response variable is the log of the average number of nests of caterpillars per tree on an area of 500 square meters.

We have \( n = 33 \) data and 10 explanatory variables.
3.8– Bayesian Variable Selection Example

- $x_1$ is the altitude (in meters),
- $x_2$ is the slope (in degrees),
- $x_3$ is the number of pines in the square,
- $x_4$ is the height (in meters) of the tree sampled at the center of the square,
- $x_5$ is the diameter of the tree sampled at the center of the square,
- $x_6$ is the index of the settlement density,
- $x_7$ is the orientation of the square (from 1 if southbound to 2 otherwise),
- $x_8$ is the height (in meters) of the dominant tree,
- $x_9$ is the number of vegetation strata,
- $x_{10}$ is the mix settlement index (from 1 if not mixed to 2 if mixed).
3.8– Bayesian Variable Selection Example
### 3.8– Bayesian Variable Selection Example

- Top five most likely models

| $\pi(\gamma|x)$ (Ridge $\delta^2 = 10$) | $\pi(\gamma|x)$ (g-p $\delta^2 = 10$) | $\pi(\gamma|x)$ (g-p, $\delta^2$ estimated) |
|--------------------------------------|--------------------------------------|--------------------------------------|
| 0,1,2,4,5/0.1946                    | 0,1,2,4,5/0.2316                      | 0,1,2,4,5/0.0929                      |
| 0,1,2,4,5,9/0.0321                   | 0,1,2,4,5,9/0.0374                    | 0,1,2,4,5,9/0.0325                    |
| 0,12,4,5,10/0.0327                   | 0,1,9/0.0344                          | 0,1,2,4,5,10/0.0295                   |
| 0,1,2,4,5,7/0.0306                   | 0,1,2,4,5,10/0.0328                   | 0,1,2,4,5,7/0.0231                    |
| 0,1,2,4,5,8/0.0251                   | 0,1,4,5/0.0306                        | 0,1,2,4,5,8/0.0228                    |
3.8– Bayesian Variable Selection Example

• This very simple sampler is much more efficient than the ones where \( \gamma \) is sampled conditional upon \((\beta, \sigma^2)\).

• However, it can also mix very slowly because the components are updated one at a time.

• It is possible to compared to true values for fixed \( \delta^2 \) and 20000 iterations appears sufficient.

• Updating correlated components together would increase significantly the convergence speed of the algorithm at the cost of an increased complexity.
4.1– Finite Mixture of Distributions

Velocity (km/sc) of galaxies in the Corona Borealis Region
4.1– Finite Mixture of Distributions

• Consider the case where one has $n$ i.i.d. data $X_i$

$$X_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \sigma_k^2)$$

where $K$ is fixed and $\theta = \{\mu_k, \sigma_k^2, p_k\}_{k=1,\ldots,K}$ are unknown.

• A standard approach consists of finding a local maximum of the log-likelihood

$$\log f(x_{1:n} | \theta) = \sum_{i=1}^{n} \log f(x_i | \theta)$$

where

$$f(x | \theta) = \sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi\sigma_k}} \exp \left( -\frac{(x - \mu_k)^2}{2\sigma_k^2} \right).$$

• Problem: The likelihood is unbounded.
4.2– Bayesian Mixture Model

• We consider the Bayesian framework where we set priors

\[
\pi(\theta) = \pi(p_1, \ldots, p_K) \prod_{k=1}^{K} \pi(\mu_k, \sigma_k^2)
\]

where

\[
(p_1, \ldots, p_K) \sim \mathcal{D}(\gamma_1, \ldots, \gamma_K).
\]

\[
\mu_k | \sigma_k^2 \sim \mathcal{N}\left(\alpha_k, \frac{\sigma_k^2}{\lambda_k}\right), \quad \sigma_k^2 \sim \mathcal{IG}\left(\frac{\lambda_k + 3}{2}, \frac{\beta_k}{2}\right).
\]

• It is impossible to use the Gibbs sampler to sample from \(\pi(\theta| x_{1:n})\).
4.2– Bayesian Mixture Model

• We can introduce the missing data $Z_i \in \{1, \ldots, K\}$ such that

$$X_i | Z_i \sim \mathcal{N} (\mu_{Z_i}, \sigma^2_{Z_i})$$

and

$$\Pr (Z_i = k) = p_k.$$

• The “complete” likelihood admits a simple form

$$\pi (x_{1:n}, z_{1:n} | \theta) = \prod_{k=1}^{n} f (x_i | \theta, z_i) \pi (z_i | \theta).$$

• Thus we propose to sample the joint posterior $\pi (\theta, z_{1:n} | y_{1:n})$ through MCMC.
4.3– Gibbs Sampler for Finite Mixture of Distributions

• We have

\[ \pi (z_{1:n} | \theta, x_{1:n}) = \prod_{i=1}^{n} \pi (z_i | \theta, x_i) \]

where

\[ \pi (z_i = j | \theta, x_i) = \frac{f(x_i | \theta, j)p_j}{\sum_{k=1}^{K} f(x_i | \theta, k)p_k} . \]

• We have

\[ \pi (\theta | z_{1:n}, x_{1:n}) = \pi (p_1, \ldots, p_K | z_{1:n}) \prod_{k=1}^{K} \pi (\mu_k, \sigma_k^2 | z_{1:n}, x_{1:n}) \]
4.3– Gibbs Sampler for Finite Mixture of Distributions

- Introducing

\[ n_k = \sum_{i=1}^{n} 1\{k\} (z_i), \quad n_k \bar{x}_k = \sum_{i=1}^{n} x_i 1\{k\} (z_i), \quad s_k^2 = \sum_{i=1}^{n} (x_i - \bar{x}_k)^2 1\{k\} (z_i). \]

- We have the full conditionals

\[ p_1, \ldots, p_K | z_{1:n} \sim D(\gamma_1 + n_1, \ldots, \gamma_K + n_K) \]

\[ \sigma_k^2 | z_{1:n}, x_{1:n} \sim IG \left( \frac{\lambda_k + n_k + 3}{2}, \frac{\lambda_k s_k^2 + \beta_k + s_k^2 - (\lambda_k + n_k)^{-1} (\lambda_k \alpha_k + n_k \bar{x}_k)^2}{2} \right) \]

\[ \mu_k | \sigma_k^2, z_{1:n}, x_{1:n} \sim N \left( \frac{\lambda_k \alpha_k + n_k \bar{x}_k}{\lambda_k + n_k}, \frac{\sigma_k^2}{\lambda_k + n_k} \right). \]

- It is thus trivial to implement the Gibbs sampler.
4.3– Gibbs Sampler for Finite Mixture of Distributions

- Consider some $n = 100$ simulated data

\[ X_i \sim 0.3N(-2, 1) + 0.7N(2, 1), \]

i.e. we have well-separated components.

- We set $\gamma_k = 1$, $\alpha_k = 0$, $\lambda_k = 0.01$, $\beta_k = 0.01$ and run the Gibbs sampler for 10000 iterations.

- We obtain $\hat{E}(\mu_1 | x_{1:n}) = 2.17$, $\hat{E}(\mu_2 | x_{1:n}) = -1.89$, $\hat{E}(\sigma_1^2 | x_{1:n}) = 0.92$, $\hat{E}(\sigma_2^2 | x_{1:n}) = 1.3$, $\hat{E}(p_1 | x_{1:n}) = 0.32$ and $\hat{E}(p_2 | x_{1:n}) = 0.68$.

- Increasing the number of iterations to 100000, I obtain similar results. Should I be happy?
• You should be extremely unhappy... as one should get

\[ E(\mu_1 | x_{1:n}) = E(\mu_2 | x_{1:n}), \ E(\sigma_1^2 | x_{1:n}) = E(\sigma_2^2 | x_{1:n}), \]

\[ E(p_1 | x_{1:n}) = E(p_2 | x_{1:n}) = 0.5. \]

• Indeed, the prior and likelihood are exchangeable and

\[
\pi \left( p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2 | x_{1:n} \right) \\
= \pi \left( p_{\zeta(1)}, \ldots, p_{\zeta(K)}, \mu_{\zeta(1)}, \ldots, \mu_{\zeta(K)}, \sigma_{\zeta(1)}^2, \ldots, \sigma_{\zeta(K)}^2 | x_{1:n} \right)
\]

for any permutation \( \zeta \) of the labels.

• Clearly, conditional expectations are not useful in this case.

⇒ This does NOT mean that your Bayesian model is useless.
4.3– Gibbs Sampler for Finite Mixture of Distributions

• One can select another point estimates; e.g. the MAP estimate

\[ \theta_{MAP} = \arg \max \pi(\theta|x_{1:n}). \]

• Alternatively, constraints can be set on the priors; e.g. we ensure that

\[ \mu_1 \leq \mu_2 \leq \ldots \leq \mu_P \]

⇒ However, this can lead to “strange” shapes of the posteriors and is not natural in most cases.

• If no constraint is ensured, then one can check whether the algorithm “mixes” by monitoring the conditional expectations.
4.3– Gibbs Sampler for Finite Mixture of Distributions

• One way to improve the algorithm consists of randomly permuting the labels (Fruwirth-Schnatter, JASA, 2002)

⇒ Realistic if $K$ is moderate because there are $K!$ permutations.

• Alternative ways to improve the algorithm include

  • Not introducing the latent variables and using sampling strategies different from Gibbs.

  • Integrating out $\theta$!
4.3– Gibbs Sampler for Finite Mixture of Distributions

- The marginal distribution of $z_{1:n}$ can be computed analytically (for conjugate priors)

$$
\pi (z_{1:n} \mid x_{1:n}) = \int \pi (z_{1:n}, \theta \mid x_{1:n}) d\theta.
$$

- $\pi (z_{1:n} \mid x_{1:n})$ is a discrete distribution with $K^n >> 1$ potential values.

- We can sample easily from it using Gibbs and using permutation moves.
Initialization:
- Select deterministically or randomly $z_{1:n}^{(0)}$.

Iteration $i; i \geq 1$:
For $k = 1 : n$
- Sample $z_k^{(i)} \sim \pi \left( z_k | x_{1:n}, z_{-k}^{(i)} \right)$.
where $z_{-k}^{(i)} = \left( z_1^{(i)}, \ldots, z_{k-1}^{(i)}, z_{k+1}^{(i-1)}, \ldots, z_n^{(i-1)} \right)$.
- Sample $\theta^{(i)} \sim \pi \left( \theta | x_{1:n}, z_{1:n}^{(i)} \right)$.

We also introduce randomly permutations of the labels.
4.3– Gibbs Sampler for Finite Mixture of Distributions

Predictive distribution for the galaxy dataset.
4.3– Gibbs Sampler for Finite Mixture of Distributions

- The Gibbs sampler is a generic tool to sample approximately from high-dimensional distributions.

- Each time you face a problem, you need to think hard about it to design an efficient algorithm.

- Except the choice of the partitions of parameters, the Gibbs sampler is parameter free; this does not mean it is efficient.