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.

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

- Initialization:

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

- Iteration i; $i \geq 1$:

  For $k = 1 : p$

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

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

2.3– Random Scan Gibbs sampler

- Initialization:
  - Select deterministically or randomly $\theta^{(0)} = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:
  - Sample $K \sim U_{\{1,...,p\}}$.
  - Set $\theta_{-K}^{(i)} = \theta_{-K}^{(i-1)}$.
  - Sample $\theta_K^{(i)} \sim \pi\left(\theta_K | \, \theta_{-K}^{(i)}\right)$.

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

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

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

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

$$\beta_k \sim \frac{1}{2}\mathcal{N}\left(0, \alpha^2\delta^2\sigma^2\right) + \frac{1}{2}\mathcal{N}\left(0, \delta^2\sigma^2\right)$$

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

$$\Pr(\gamma_k = 0) = \Pr(\gamma_k = 1) = \tfrac{1}{2},$$

$$\beta_k | \gamma_k = 0 \sim \mathcal{N}\left(0, \alpha^2\delta^2\sigma^2\right), \quad \beta_k | \gamma_k = 1 \sim \mathcal{N}\left(0, \delta^2\sigma^2\right).$$

3.2– A Bad Gibbs Sampler

- We have parameters $\left(\beta_{1:p}, \gamma_{1:p}, \sigma^2\right)$ and observe n observations $D = \{x_j, y_j\}_{j=1}^{n}$.

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

- In particular

$$p\left(\gamma_{1:p} \middle| D, \beta_{1:p}, \sigma^2\right) = \prod_{k=1}^{p} p\left(\gamma_k \middle| \beta_k, \sigma^2\right)$$

and

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

- The Gibbs sampler becomes reducible as α goes to zero.

3.3– Bayesian Variable Selection Example

• This is the result of bad modelling and bad algorithm.

You would like to put $\alpha \simeq 0$ 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 β_k is defined even when $\gamma_k = 0$.

• A neater way to write such models is to write

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

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

• Prior distributions

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

and $\pi(\gamma) = \prod_{k=1}^{p} \pi(\gamma_k) = 2^{-p}$.

- We are interested in sampling from the trans-dimensional distribution $\pi\left(\gamma, \beta_\gamma, \sigma^2 \middle| D\right)$

- However, we know that

$$\pi\left(\gamma, \beta_\gamma, \sigma^2 \middle| D\right) = \pi\left(\gamma | D\right) \pi\left(\beta_\gamma, \sigma^2 \middle| D, \gamma\right)$$

where

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

and

$$\begin{aligned} \pi\left(D | \gamma\right) &= \int \pi\left(D, \beta_\gamma, \sigma^2 \middle| \gamma\right) d\beta_\gamma d\sigma^2 \\ &\propto \Gamma\left(\frac{\nu_0 + n}{2} + 1\right) \delta^{-n_\gamma} \left|\Sigma_\gamma\right|^{1/2} \left(\frac{\gamma_0 + \sum_{j=1}^{n} y_k^2 - \mu_\gamma^{\mathrm{T}} \Sigma_\gamma^{-1} \mu_\gamma}{2}\right)^{-\left(\frac{\nu_0+n}{2}+1\right)} . \end{aligned}$$

- The full conditional distribution for $\pi\left(\beta_\gamma, \sigma^2 \middle| D, \gamma\right)$ is

$$\begin{aligned}\pi_\gamma\left(\beta_\gamma, \sigma^2 \middle| D\right) &= \mathcal{N}\left(\beta_\gamma; \mu_\gamma, \sigma^2\Sigma_\gamma\right) \\ &\quad \times \mathcal{IG}\left(\sigma^2; \frac{\nu_0 + n}{2}, \frac{\gamma_0 + \sum_{j=1}^n y_j^2 - \mu_\gamma^{\mathrm{T}}\Sigma_\gamma^{-1}\mu_\gamma}{2}\right)\end{aligned}$$

where

$$\mu_\gamma = \Sigma_\gamma\left(\sum_{j=1}^n y_j x_{\gamma,j}\right), \; \Sigma_\gamma^{-1} = \delta^{-2} I_{n_\gamma} + \sum_{i=1}^n x_{\gamma,j} x_{\gamma,j}^{\mathrm{T}}.$$

• 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)

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

• Robust models where additionally one has

$$\delta^2 \sim \mathcal{IG}\left(\frac{a_0}{2}, \frac{b_0}{2}\right).$$

• Such variations are very important and can modify dramatically the performance of the Bayesian model.

3.5– Collapsed Gibbs Sampler for Bayesian Variable Selection

- $\pi\left(\gamma|\, D\right)$ is a discrete probability distribution with 2^p potential values, we assume δ^2 *known* here.
- Initialization:
  - Select deterministically or randomly $\gamma^{(0)} = \left(\gamma_1^{(0)}, ..., \gamma_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:

  For $k = 1 : p$

  - Sample $\gamma_k^{(i)} \sim \pi\left(\gamma_k|\, D, \gamma_{-k}^{(i)}\right)$.

  where $\gamma_{-k}^{(i)} = \left(\gamma_1^{(i)}, ..., \gamma_{k-1}^{(i)}, \gamma_{k+1}^{(i-1)}, ..., \gamma_p^{(i-1)}\right)$.

  - Optional step: Sample $\left(\beta_\gamma^{(i)}, \sigma^{2(i)}\right) \sim \pi\left(\beta_\gamma, \sigma^2 \middle|\, D, \gamma^{(i)}\right)$.

- Consider the case where δ^2 is unknown.
- Initialization:
  - Select deterministically or randomly $\left(\gamma^{(0)}, \beta_\gamma^{(0)}, \sigma^{2(0)}, \delta^{2(0)}\right)$
- Iteration i; $i \geq 1$:

  For $k = 1 : p$

  - Sample $\gamma_k^{(i)} \sim \pi\left(\gamma_k \middle| D, \gamma_{-k}^{(i)}, \delta^{2(i-1)}\right)$.

  where $\gamma_{-k}^{(i)} = \left(\gamma_1^{(i)}, ..., \gamma_{k-1}^{(i)}, \gamma_{k+1}^{(i-1)}, ..., \gamma_p^{(i-1)}\right)$.

  - Sample $\left(\beta_\gamma^{(i)}, \sigma^{2(i)}\right) \sim \pi\left(\beta_\gamma, \sigma^2 \middle| D, \gamma^{(i)}, \delta^{2(i)}\right)$.
  - Sample $\delta^{2(i)} \sim \pi\left(\delta^{2(i)} \middle| \beta_\gamma^{(i)}\right)$.

3.7– Bayesian Variable Selection Example

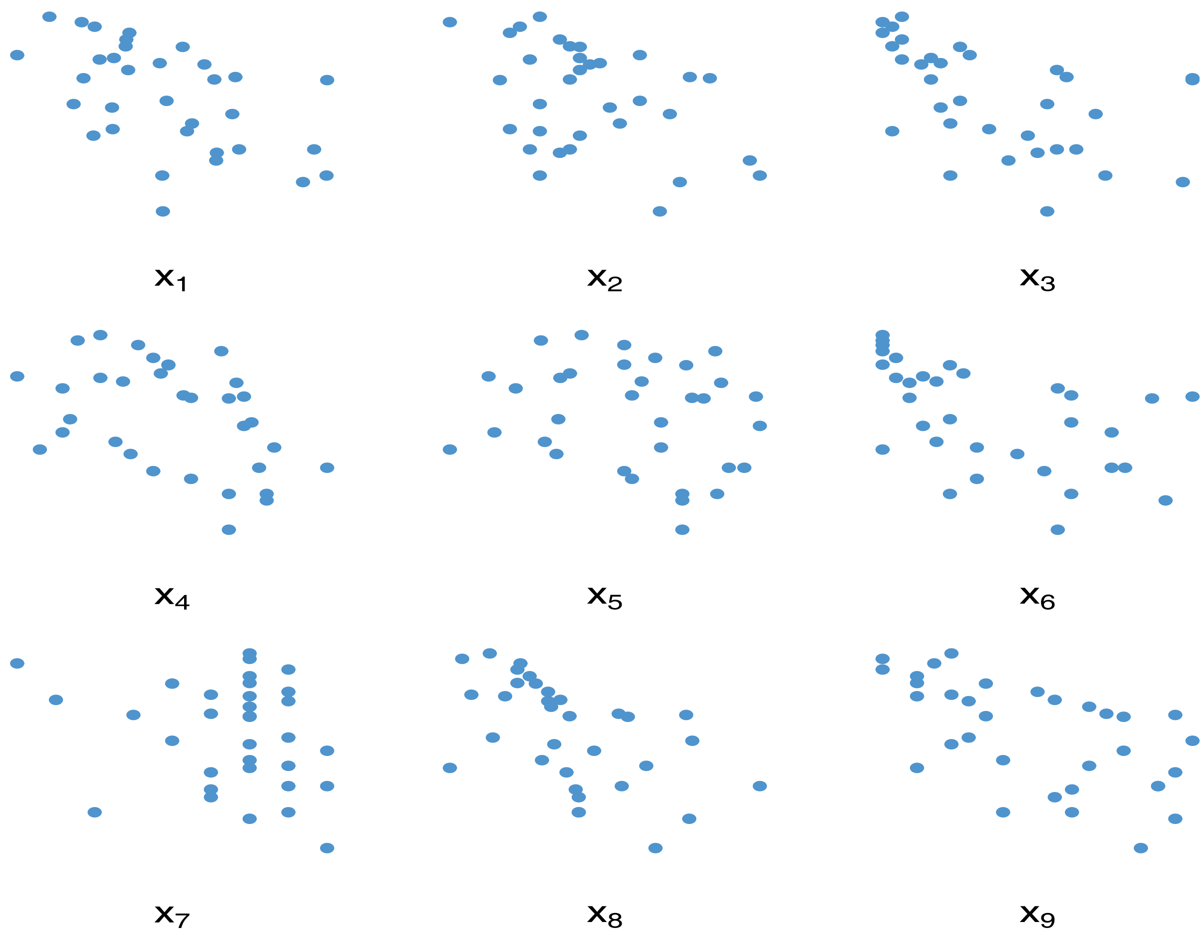
• Caterpillar dataset: 1973 study to assess the influence of some forest settlement characteristics on the development of catepillar colonies.

• 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

- 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



- Top five most likely models

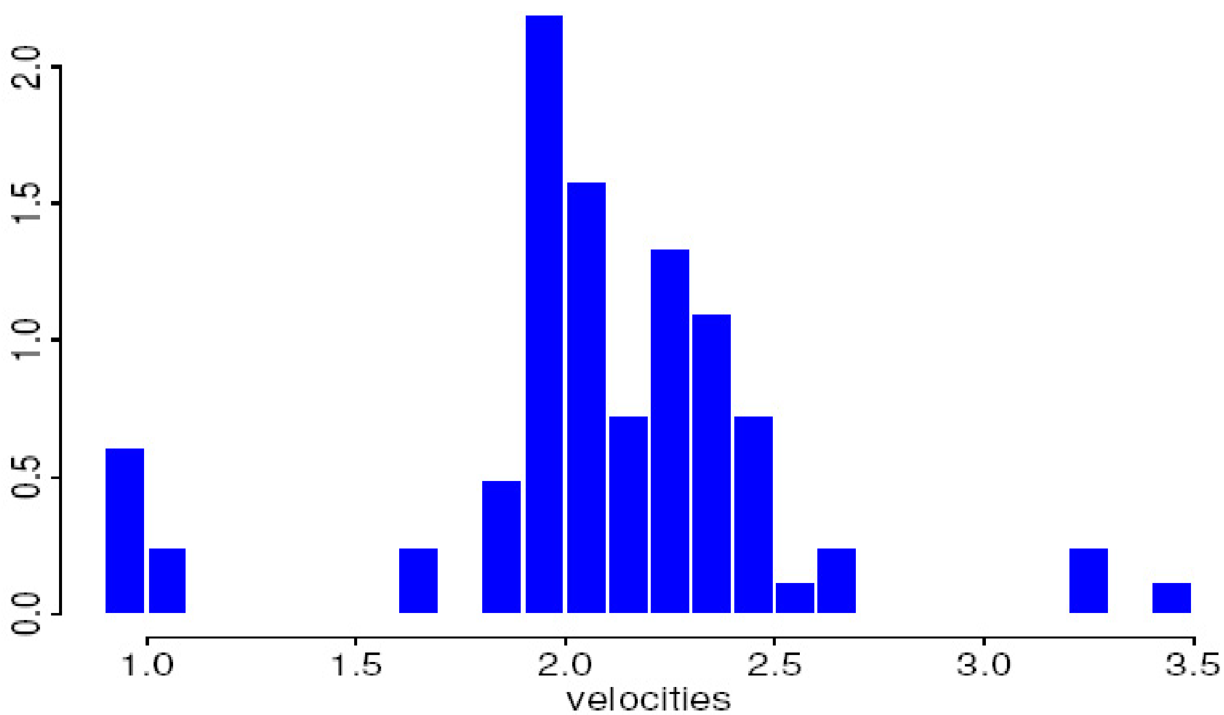
$\pi(\gamma\vert x)$ (Ridge $\delta^2 = 10$)	$\pi(\gamma\vert x)$ (g-p $\delta^2 = 10$)	$\pi(\gamma\vert x)$ (g-p, δ^2 estimated)
0,1,2,4,5/0.1946	0,1,2,4,5/0.2316	0,1,2,4,5/0.0929
0,1,2,4,5,9/0.0321	0,1,2,4,5,9/0.0374	0,1,2,4,5,9/0.0325
0,12,4,5,10/0.0327	0,1,9/0.0344	0,1,2,4,5,10/0.0295
0,1,2,4,5,7/0.0306	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

- This very simple sampler is much more efficient than the ones where γ is sampled conditional upon $\left(\beta, \sigma^2\right)$.

- 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 δ^2 and 20000 iterations appears sufficients.

- Updating correlated components together would increase significantly the convergence speed of the algorithm at the cost of an increased complexity.



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

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

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

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

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

where

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

- Problem: The likelihood is unbounded.

4.2– Bayesian Mixture Model

- We consider the Bayesian framework where we set priors

$$\pi(\theta) = \pi(p_1, ..., p_K) \prod_{k=1}^{K} \pi\left(\mu_k, \sigma_k^2\right)$$

where

$$\begin{aligned} (p_1, ..., p_K) &\sim \mathcal{D}(\gamma_1, ..., \gamma_K). \\ \mu_k | \sigma_k^2 &\sim \mathcal{N}\left(\alpha_k, \frac{\sigma_k^2}{\lambda_k}\right), \ \sigma_k^2 \sim \mathcal{IG}\left(\frac{\lambda_k + 3}{2}, \frac{\beta_k}{2}\right). \end{aligned}$$

- 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 $Z_i \in \{1, ..., K\}$ such that

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

and

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

- The "complete" likelihood admits a simple form

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

- Thus we propose to sample the joint posterior $\pi\left(\theta, z_{1:n} | \, y_{1:n}\right)$ through MCMC.

4.3– Gibbs Sampler for Finite Mixture of Distributions

- We have

$$\pi\left(z_{1:n}\middle|\,\theta,x_{1:n}\right)=\prod_{i=1}^{n}\pi\left(z_i\middle|\,\theta,x_i\right)$$

where

$$\pi\left(z_i=j\middle|\,\theta,x_i\right)=\frac{f\left(x_i\middle|\,\theta,j\right)p_j}{\sum_{k=1}^{K}f\left(x_i\middle|\,\theta,k\right)p_k}.$$

- We have

$$\pi\left(\theta\middle|\,z_{1:n},x_{1:n}\right)=\pi\left(p_1,...,p_K\middle|\,z_{1:n}\right)\prod_{k=1}^{K}\pi\left(\mu_k,\sigma_k^2\middle|\,z_{1:n},x_{1:n}\right)$$

- Introducing

$$n_k = \sum_{i=1}^{n} \mathbf{1}_{\{k\}}(z_i), \; n_k \overline{x}_k = \sum_{i=1}^{n} x_i \mathbf{1}_{\{k\}}(z_i), \; s_k^2 = \sum_{i=1}^{n} (x_i - \overline{x}_k)^2 \mathbf{1}_{\{k\}}(z_i).$$

- We have the full conditionals

$$p_1, ..., p_K | z_{1:n} \sim \mathcal{D}(\gamma_1 + n_1, ..., \gamma_K + n_K)$$

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

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

- It is thus trivial to implement the Gibbs sampler.

• Consider some $n = 100$ simulated data

$$X_i \sim 0.3\mathcal{N}(-2,1) + 0.7\mathcal{N}(2,1),$$

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

• Increasing the number of iterations to 100000, I obtain similar results. Should I be happy?

4.3– Gibbs Sampler for Finite Mixture of Distributions

- You should be extremely unhappy... as one should get

$$
\begin{aligned}
E\left(\mu_1 \middle| x_{1:n}\right) &= E\left(\mu_2 \middle| x_{1:n}\right),\ E\left(\sigma_1^2 \middle| x_{1:n}\right) = E\left(\sigma_2^2 \middle| x_{1:n}\right), \\
E\left(p_1 \middle| x_{1:n}\right) &= E\left(p_2 \middle| x_{1:n}\right) = 0.5.
\end{aligned}
$$

- Indeed, the prior and likelihood are exchangeable and

$$
\begin{aligned}
& \pi\left(p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2 \middle| x_{1:n}\right) \\
= & \pi\left(p_{\zeta(1)}, ..., p_{\zeta(K)}, \mu_{\zeta(1)}, ..., \mu_{\zeta(K)}, \sigma_{\zeta(1)}^2, ..., \sigma_{\zeta(K)}^2 \middle| x_{1:n}\right)
\end{aligned}
$$

for any permutation ζ of the labels.

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

$\Rightarrow$ 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

$$\theta_{MAP} = \arg\max \pi \left(\theta | \, x_{1:n} \right).$$

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

$$\mu_1 \leq \mu_2 \leq ... \leq \mu_P$$

$\Rightarrow$ 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 θ!

4.3– Gibbs Sampler for Finite Mixture of Distributions

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

$$\pi\left(z_{1:n}\middle|\, x_{1:n}\right) = \int \pi\left(z_{1:n}, \theta\middle|\, x_{1:n}\right) d\theta.$$

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

- We can sample easily from it using Gibbs and using permutation moves.

4.3– Gibbs Sampler for Finite Mixture of Distributions

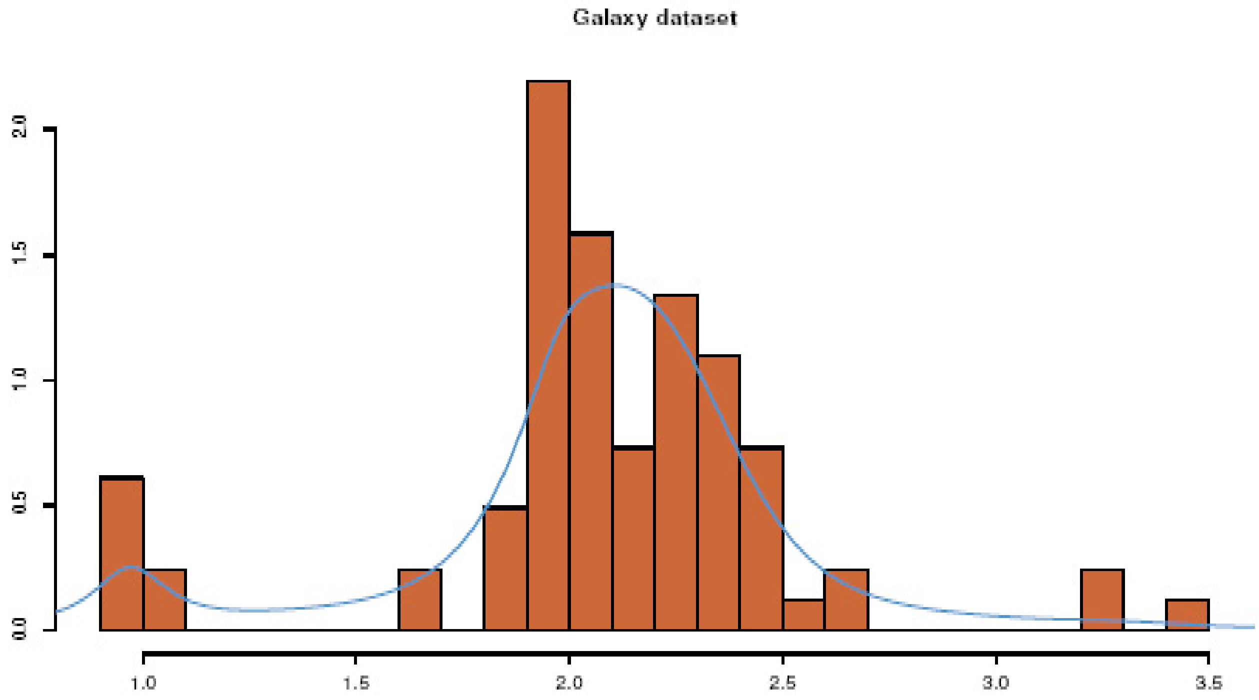
- Initialization:
  - Select deterministically or randomly $z_{1:n}^{(0)}$.
- Iteration i; $i \geq 1$:

  For $k = 1 : n$
  - Sample $z_k^{(i)} \sim \pi\left(z_k \middle| \, x_{1:n}, z_{-k}^{(i)} \right)$.

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

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.