Advanced Simulation - Lecture 6

Patrick Rebeschini

January 31th, 2018

Markov chain Monte Carlo

- We are interested in sampling from a distribution π, for instance a posterior distribution in a Bayesian framework.

- Markov chains with π as invariant distribution can be constructed to approximate expectations with respect to π.

- For example, the Gibbs sampler generates a Markov chain targeting π defined on $\mathbb{R}^d$ using the full conditionals

$$\pi(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d).$$

Gibbs Sampling

- Assume you are interested in sampling from

$$\pi\left(x\right)=\pi\left(x_1,x_2,...,x_d\right).$$

- Notation: $x_{-i}:=\left(x_1,...,x_{i-1},x_{i+1},...,x_d\right)$.

Systematic scan Gibbs sampler. Let $\left(X_1^{(1)},...,X_d^{(1)}\right)$ be the initial state then iterate for $t=2,3,...$

1. Sample $X_1^{(t)}\sim\pi_{X_1|X_{-1}}\left(\cdot|\,X_2^{(t-1)},...,X_d^{(t-1)}\right)$.
   ...

j. Sample
   $X_j^{(t)}\sim\pi_{X_j|X_{-j}}\left(\cdot|\,X_1^{(t)},...,X_{j-1}^{(t)},X_{j+1}^{(t-1)},...,X_d^{(t-1)}\right)$.

...

d. Sample $X_d^{(t)}\sim\pi_{X_d|X_{-d}}\left(\cdot|\,X_1^{(t)},...,X_{d-1}^{(t)}\right)$.

Gibbs Sampling

- Is the joint distribution π uniquely specified by the conditional distributions $\pi_{X_i|X_{-i}}$?

- Does the Gibbs sampler provide a Markov chain with the correct stationary distribution π?

- If yes, does the Markov chain converge towards this invariant distribution?

- It will turn out to be the case under some mild conditions.

Hammersley-Clifford Theorem

- **Theorem**. Consider a distribution whose density $\pi(x_1, x_2, ..., x_d)$ is such that $\text{supp}(\pi) = \otimes_{i=1}^{d}\text{supp}(\pi_{X_i})$. Then for any $(z_1, ..., z_d) \in \text{supp}(\pi)$, we have

$$\pi(x_1, x_2, ..., x_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}(x_j|\, x_{1:j-1}, z_{j+1:d})}{\pi_{X_j|X_{-j}}(z_j|\, x_{1:j-1}, z_{j+1:d})}.$$

- Proof: we have

$$\begin{aligned} \pi(x_{1:d-1}, x_d) &= \pi_{X_d|X_{-d}}(x_d|\, x_{1:d-1})\pi(x_{1:d-1}), \\ \pi(x_{1:d-1}, z_d) &= \pi_{X_d|X_{-d}}(z_d|\, x_{1:d-1})\pi(x_{1:d-1}). \end{aligned}$$

Therefore

$$\pi(x_{1:d}) = \pi(x_{1:d-1}, z_d)\frac{\pi_{X_d|X_{-d}}(x_d|\, x_{1:d-1})}{\pi_{X_d|X_{-d}}(z_d|\, x_{1:d-1})}.$$

Hammersley-Clifford Theorem

- Similarly, we have

$$\pi\left(x_{1:d-1}, z_d\right) = \pi_{X_{d-1}|X_{-(d-1)}}\left(x_{d-1}\middle| x_{1:d-2}, z_d\right)\pi\left(x_{1:d-2}, z_d\right),$$
$$\pi\left(x_{1:d-2}, z_{d-1}, z_d\right) = \pi_{X_{d-1}|X_{-(d-1)}}\left(z_{d-1}\middle| x_{1:d-2}, z_d\right)\pi\left(x_{1:d-2}, z_d\right)$$

hence

$$\pi\left(x_{1:d}\right) = \pi(x_{1:d-2}, z_{d-1}, z_d)\frac{\pi_{X_{d-1}|X_{-(d-1)}}\left(x_{d-1}\middle| x_{1:d-2}, z_d\right)}{\pi_{X_{d-1}|X_{-(d-1)}}\left(z_{d-1}\middle| x_{1:d-2}, z_d\right)} \times \frac{\pi_{X_d|X_{-d}}\left(x_d\middle| x_{1:d-1}\right)}{\pi_{X_d|X_{-d}}\left(z_d\middle| x_{1:d-1}\right)}$$

- By iterating, we obtain the theorem, where the multiplicative constant is exactly $\pi(z_1, \ldots, z_d)$.

Example: Non-Integrable Target

- Consider the following conditionals on $\mathbb{R}^+$

$$\pi_{X_1|X_2}(x_1|\,x_2) = x_2 \exp(-x_2 x_1)$$
$$\pi_{X_2|X_1}(x_2|\,x_1) = x_1 \exp(-x_1 x_2).$$

  We might expect that these full conditionals define a joint probability density $\pi(x_1, x_2)$.

- Hammersley-Clifford would give

$$\pi(x_1, x_2, ..., x_d) \propto \frac{\pi_{X_1|X_2}(x_1|\,z_2)}{\pi_{X_1|X_2}(z_1|\,z_2)} \frac{\pi_{X_2|X_1}(x_2|\,x_1)}{\pi_{X_2|X_1}(z_2|\,x_1)}$$
$$= \frac{z_2 \exp(-z_2 x_1)\, x_1 \exp(-x_1 x_2)}{z_2 \exp(-z_2 z_1)\, x_1 \exp(-x_1 z_2)} \propto \exp(-x_1 x_2).$$

  However $\int\int \exp(-x_1 x_2)\, dx_1 dx_2$ is not finite so $\pi_{X_1|X_2}(x_1|\,x_2) = x_2 \exp(-x_2 x_1)$ and $\pi_{X_2|X_1}(x_1|\,x_2) = x_1 \exp(-x_1 x_2)$ are not compatible.

Example: Positivity condition violated

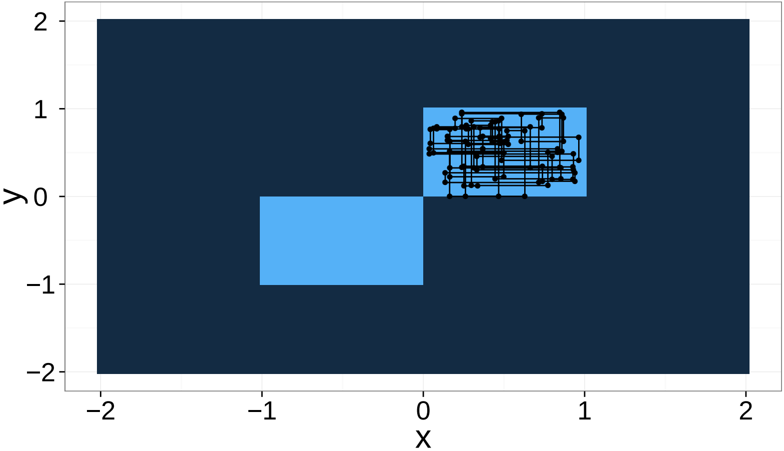


Figure: Gibbs sampling targeting
$\pi(x, y) \propto \mathbb{1}_{[-1,0]\times[-1,0]\cup[0,1]\times[0,1]}(x, y)$.

Invariance of the Gibbs sampler

- The kernel of the Gibbs sampler (case $d = 2$) is

$$K(x^{(t-1)}, x^{(t)}) = \pi_{X_1|X_2}(x_1^{(t)} \mid x_2^{(t-1)}) \pi_{X_2|X_1}(x_2^{(t)} \mid x_1^{(t)})$$

- Case $d > 2$:

$$K(x^{(t-1)}, x^{(t)}) = \prod_{j=1}^{d} \pi_{X_j|X_{-j}}(x_j^{(t)} \mid x_{1:j-1}^{(t)}, x_{j+1:d}^{(t-1)})$$

- **Proposition**: The systematic scan Gibbs sampler kernel admits π as invariant distribution.
- *Proof for $d = 2$.* We have

$$\int K(x,y)\pi(x)dx = \int \pi(y_2 \mid y_1)\pi(y_1 \mid x_2)\pi(x_1, x_2)dx_1 dx_2$$
$$= \pi(y_2 \mid y_1) \int \pi(y_1 \mid x_2)\pi(x_2)dx_2$$
$$= \pi(y_2 \mid y_1)\pi(y_1) = \pi(y_1, y_2) = \pi(y).$$

Irreducibility and Recurrence

- **Proposition**: Assume π satisfies the positivity condition, then the Gibbs sampler yields a $\pi-$irreducible and recurrent Markov chain.

- **Theorem**. Assume the positivity condition is satisfied then we have for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

$$\lim \frac{1}{t} \sum_{i=1}^{t} \varphi\left(X^{(i)}\right) = \int_{\mathbb{X}} \varphi\left(x\right) \pi\left(x\right) dx$$

  for $\pi-$almost all starting value $X^{(1)}$.

Example: Bivariate Normal Distribution

- Let $X := (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2)$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}.$$

- The Gibbs sampler proceeds as follows in this case

1. Sample $X_1^{(t)} \sim \mathcal{N}\left(\mu_1 + \rho/\sigma_2^2 \left(X_2^{(t-1)} - \mu_2\right), \sigma_1^2 - \rho^2/\sigma_2^2\right)$
2. Sample $X_2^{(t)} \sim \mathcal{N}\left(\mu_2 + \rho/\sigma_1^2 \left(X_1^{(t)} - \mu_1\right), \sigma_2^2 - \rho^2/\sigma_1^2\right)$.

- By proceeding this way, we generate a Markov chain $X^{(t)}$ whose successive samples are correlated. If successive values of $X^{(t)}$ are strongly correlated, then we say that the Markov chain mixes slowly.

Bivariate Normal Distribution

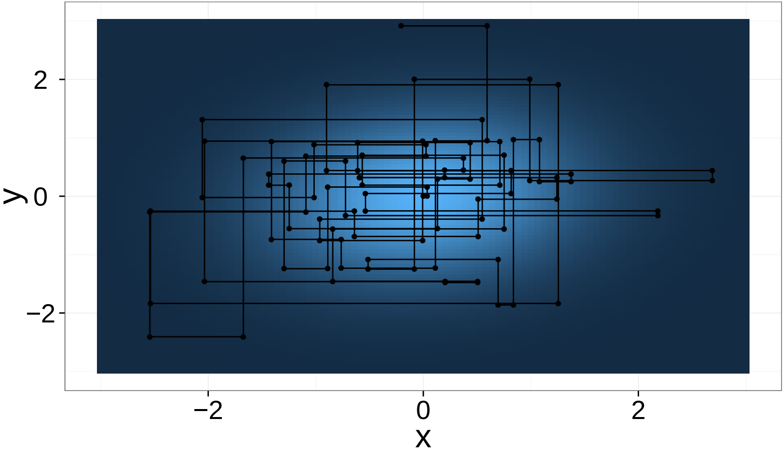


Figure: Case where $\rho = 0.1$, first 100 steps.

Bivariate Normal Distribution

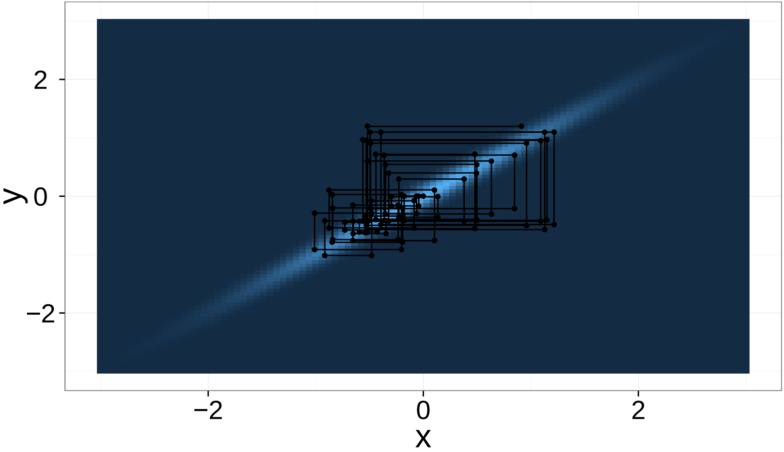


Figure: Case where $\rho = 0.99$, first 100 steps.

Bivariate Normal Distribution

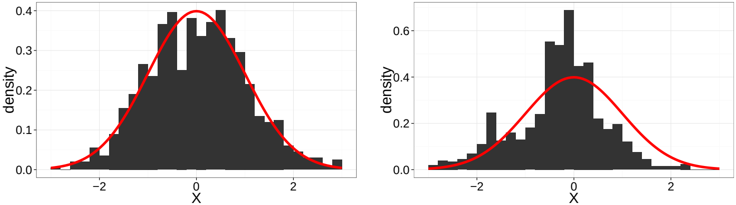


Figure: Histogram of the first component of the chain after 1000 iterations. Small ρ on the left, large ρ on the right.

Bivariate Normal Distribution

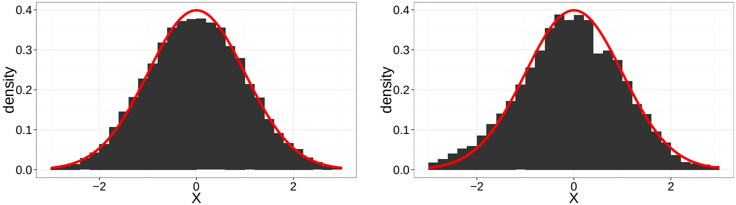


Figure: Histogram of the first component of the chain after 10000 iterations. Small ρ on the left, large ρ on the right.

Bivariate Normal Distribution

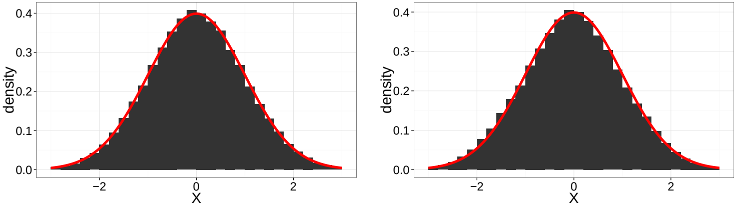


Figure: Histogram of the first component of the chain after 100000 iterations. Small ρ on the left, large ρ on the right.

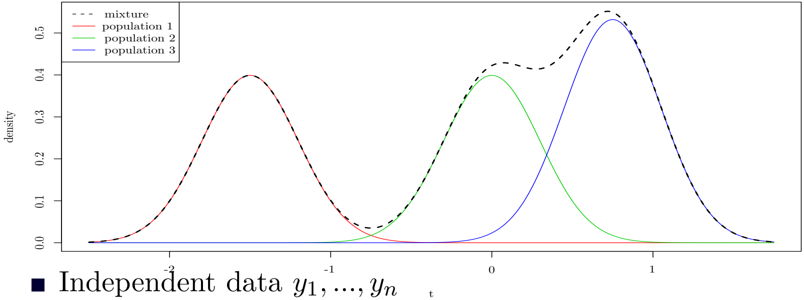
Gibbs Sampling and Auxiliary Variables

- Gibbs sampling requires sampling from $\pi_{X_j|X_{-j}}$.
- In many scenarios, we can include a set of auxiliary variables $Z_1, ..., Z_p$ and have an "extended" distribution of joint density $\overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)$ such that

$$\int \overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)\, dz_1 ... dz_d = \pi(x_1, ..., x_d).$$

which is such that its full conditionals are easy to sample.
- Mixture models, Capture-recapture models, Tobit models, Probit models etc.

Mixtures of Normals



- Independent data $y_1, ..., y_n$

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

where $\theta = (p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2)$.

Bayesian Model

- Likelihood function

$$p\left(y_1,...,y_n \middle| \theta\right) = \prod_{i=1}^{n} p\left(y_i \middle| \theta\right) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{\left(y_i - \mu_k\right)^2}{2\sigma_k^2} \right) \right).$$

  Let's fix $K = 2$, $\sigma_k^2 = 1$ and $p_k = 1/K$ for all k.

- Prior model

$$p\left(\theta\right) = \prod_{k=1}^{K} p\left(\mu_k\right)$$

  where

$$\mu_k \sim \mathcal{N}\left(\alpha_k, \beta_k\right).$$

  Let us fix $\alpha_k = 0, \beta_k = 1$ for all k.

- Not obvious how to sample $p(\mu_1 \mid \mu_2, y_1, \ldots, y_n)$.

Auxiliary Variables for Mixture Models

- Associate to each Y_i an auxiliary variable $Z_i \in \{1, ..., K\}$ such that

$$\mathbb{P}\left(Z_i = k \middle| \theta\right) = p_k \text{ and } \left. Y_i \middle| Z_i = k, \theta \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right)\right.$$

so that

$$p\left(y_i \middle| \theta\right) = \sum_{k=1}^{K} \mathbb{P}\left(Z_i = k\right) \mathcal{N}\left(y_i; \mu_k, \sigma_k^2\right)$$

- The extended posterior is given by

$$p\left(\theta, z_1, ..., z_n \middle| y_1, ..., y_n\right) \propto p\left(\theta\right) \prod_{i=1}^{n} \mathbb{P}\left(z_i \middle| \theta\right) p\left(y_i \middle| z_i, \theta\right).$$

- Gibbs samples alternately

$$\mathbb{P}\left(z_{1:n} \middle| y_{1:n}, \mu_{1:K}\right)$$
$$p\left(\mu_{1:K} \middle| y_{1:n}, z_{1:n}\right).$$

Gibbs Sampling for Mixture Model

- We have
$$\mathbb{P}(z_{1:n} | y_{1:n}, \theta) = \prod_{i=1}^{n} \mathbb{P}(z_i | y_i, \theta)$$
where
$$\mathbb{P}(z_i | y_i, \theta) = \frac{\mathbb{P}(z_i | \theta)\, p(y_i | z_i, \theta)}{\sum_{k=1}^{K} \mathbb{P}(z_i = k | \theta)\, p(y_i | z_i = k, \theta)}$$
- Let $n_k = \sum_{i=1}^{n} \mathbf{1}_{\{k\}}(z_i), n_k \overline{y}_k = \sum_{i=1}^{n} y_i \mathbf{1}_{\{k\}}(z_i)$ then
$$\mu_k | z_{1:n}, y_{1:n} \sim \mathcal{N}\left(\frac{n_k \overline{y}_k}{1+n_k}, \frac{1}{1+n_k}\right).$$

Mixtures of Normals

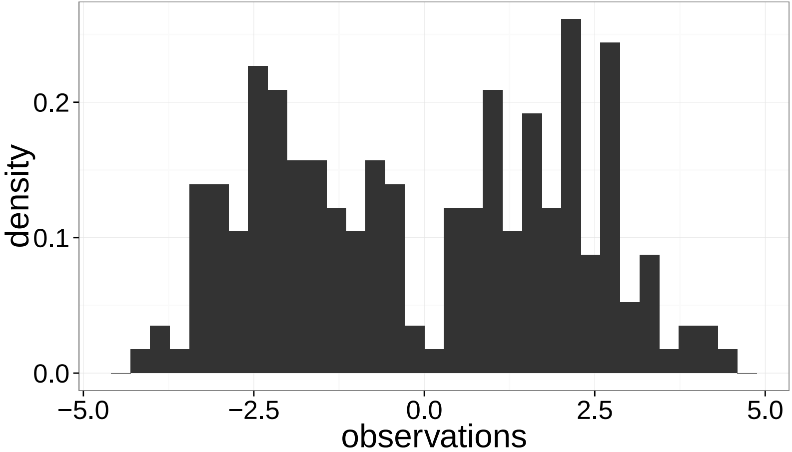


Figure: 200 points sampled from $\frac{1}{2}\mathcal{N}(-2,1) + \frac{1}{2}\mathcal{N}(2,1)$.

Mixtures of Normals

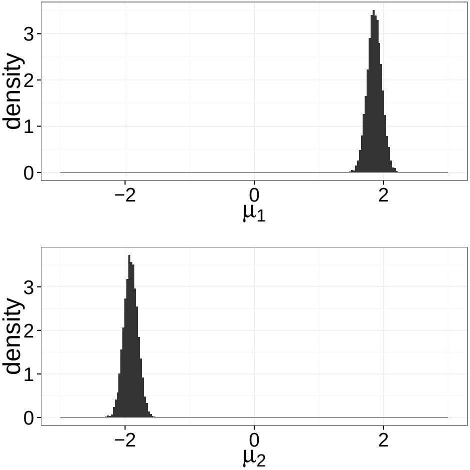


Figure: Histogram of the parameters obtained by $10,000$ iterations of Gibbs sampling.

Mixtures of Normals

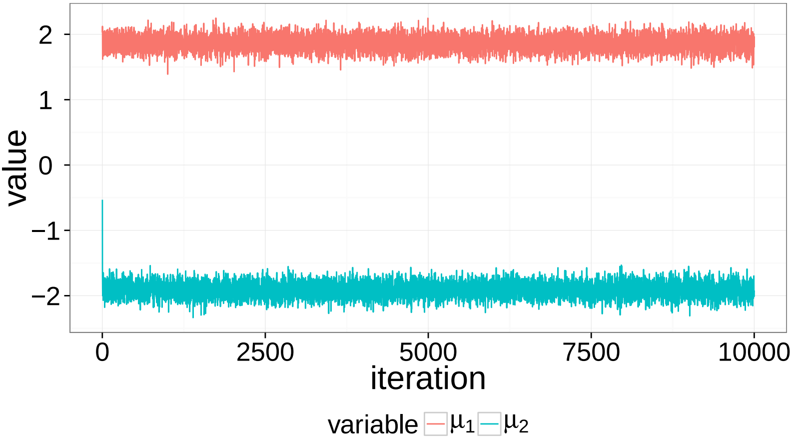


Figure: Traceplot of the parameters obtained by $10,000$ iterations of Gibbs sampling.

Gibbs sampling in practice

- Many posterior distributions can be automatically decomposed into conditional distributions by computer programs.

- This is the idea behind `BUGS` (Bayesian inference Using Gibbs Sampling), `JAGS` (Just another Gibbs Sampler).