Advanced Simulation - Lecture 5

George Deligiannidis

February 1st, 2016

Irreducibility and aperiodicity

Definition

Given a distribution μ over $\mathbb{X}$, a Markov chain is μ-irreducible if

$$\forall x \in \mathbb{X} \quad \forall A : \mu(A) > 0 \quad \exists t \in \mathbb{N} \quad K^t(x, A) > 0.$$

A μ-irreducible Markov chain of transition kernel K is periodic if there exists some partition of the state space $\mathbb{X}_1, ..., \mathbb{X}_d$ for $d \geq 2$, such that

$$\forall i, j, t, s : \ \mathbb{P}\left(X_{t+s} \in \mathbb{X}_j \middle| X_t \in \mathbb{X}_i \right) = \begin{cases} 1 & j = i + s \bmod d \\ 0 & \text{otherwise.} \end{cases} \quad .$$

Otherwise the chain is aperiodic.

Recurrence and Harris Recurrence

For any measurable set A of $\mathbb{X}$, let

$$\eta_A = \sum_{k=1}^{\infty} \mathbb{I}_A\left(X_k\right) = \text{\# of visits to } A.$$

Definition

A μ-irreducible Markov chain is recurrent if for any measurable set $A \subset \mathbb{X} : \mu\left(A\right) > 0$, then

$$\forall x \in A \quad \mathbb{E}_x\left(\eta_A\right) = \infty.$$

A μ-irreducible Markov chain is Harris recurrent if for any measurable set $A \subset \mathbb{X} : \mu\left(A\right) > 0$, then

$$\forall x \in \mathbb{X} \quad \mathbb{P}_x\left(\eta_A = \infty\right) = 1.$$

Harris recurrence is stronger than recurrence.

Invariant Distribution and Reversibility

Definition

A distribution of density π is invariant or *stationary* for a Markov kernel K, if

$$\int_{\mathbb{X}} \pi(x) K(x,y)\, dx = \pi(y).$$

A Markov kernel K is π-reversible if

$$\forall f \quad \iint f(x,y)\pi(x) K(x,y)\, dxdy = \iint f(y,x)\pi(x) K(x,y)\, dxdy$$

where f is a bounded measurable function.

Detailed balance

In practice it is easier to check the detailed balance condition:

$$\forall x, y \in \mathbb{X} \quad \pi(x)K(x,y) = \pi(y)K(y,x)$$

Lemma

If detailed balance holds, then π is invariant for K and K is π-reversible.

Example: the Gaussian AR process is π-reversible, π-invariant for

$$\pi(x) = \mathcal{N}\left(x; 0, \frac{\tau^2}{1-\rho^2}\right)$$

when $|\rho| < 1$.

Checking for recurrence

It's often straightforward to check for irreducibility, or for an invariant measure but not so for recurrence.

Proposition

If the chain is μ-irreducible and admits an invariant measure then the chain is recurrent.

Remark: A chain that is μ-irreducible and admits an invariant measure is called a positive.

Law of Large Numbers

Theorem

If K is a π-irreducible, π-invariant Markov kernel, then for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

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

almost surely, for $\pi-$ almost all starting values x.

Theorem

If K is a π-irreducible, π-invariant, Harris recurrent Markov chain, then for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

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

almost surely, for any starting value x.

Convergence

Theorem

Suppose the kernel K is π-irreducible, π-invariant, aperiodic. Then, we have

$$\lim_{t \to \infty} \int_{\mathbb{X}} \left| K^t(x, y) - \pi(y) \right| dy = 0$$

for π−almost all starting values x.

Under some additional conditions, one can prove that a chain is geometrically ergodic, i.e. there exists $\rho < 1$ and a function $M : \mathbb{X} \to \mathbb{R}^+$ such that for all measurable set A:

$$|K^n(x, A) - \pi(A)| \leq M(x)\rho^n,$$

for all $n \in \mathbb{N}$. In other words, we can obtain a rate of convergence.

Central Limit Theorem

Theorem

Under regularity conditions, for a Harris recurrent, π-invariant Markov chain, we can prove

$$\sqrt{t}\left[\frac{1}{t}\sum_{i=1}^{t}\varphi\left(X_i\right)-\int_{\mathbb{X}}\varphi\left(x\right)\pi\left(x\right)dx\right]\xrightarrow[t\to\infty]{D}\mathcal{N}\left(0,\sigma^2\left(\varphi\right)\right),$$

where the asymptotic variance can be written

$$\sigma^2\left(\varphi\right)=\mathbb{V}_\pi\left[\varphi\left(X_1\right)\right]+2\sum_{k=2}^{\infty}\mathbb{C}ov_\pi\left[\varphi\left(X_1\right),\varphi\left(X_k\right)\right].$$

This formula shows that (positive) correlations increase the asymptotic variance, compared to i.i.d. samples for which the variance would be $\mathbb{V}_\pi(\varphi(X))$.

Central Limit Theorem

- Example: for the AR Gaussian model, $\pi(x) = \mathcal{N}\left(x; 0, \tau^2/(1-\rho^2)\right)$ for $|\rho| < 1$ and

$$\mathbb{C}\text{ov}\left(X_1, X_k\right) = \rho^{k-1}\mathbb{V}\left[X_1\right] = \rho^{k-1}\frac{\tau^2}{1-\rho^2}.$$

- Therefore with $\varphi(x) = x$,

$$\sigma^2(\varphi) = \frac{\tau^2}{1-\rho^2}\left(1 + 2\sum_{k=1}^{\infty}\rho^k\right) = \frac{\tau^2}{1-\rho^2}\frac{1+\rho}{1-\rho} = \frac{\tau^2}{(1-\rho)^2},$$

which increases when $\rho \to 1$.

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

$$\pi(x) = \pi(x_1, x_2, ..., x_d), \quad x\mathbb{R}^d.$$

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

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 I

Theorem

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

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

Remark: The condition above is the positivity condition. Equivalently, if $\pi_{X_i}(x_i) > 0$ for $i = 1, \ldots, d$, then

$$\pi(x_1, \ldots, x_d) > 0.$$

Proof of Hammersley-Clifford Theorem

Proof.

We have

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

Therefore

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

Proof.

Similarly, we have

$$\pi(x_{1:d-1}, z_d) = \pi(x_{1:d-2}, z_{d-1}, z_d) \frac{\pi(x_{1:d-1}, z_d)}{\pi(x_{1:d-2}, z_{d-1}, z_d)}$$
$$= \pi(x_{1:d-2}, z_{d-1}, z_d) \frac{\pi(x_{1:d-1}, z_d) / \pi(x_{1:d-2}, z_d)}{\pi(x_{1:d-2}, z_{d-1}, z_d) / \pi(x_{1:d-2}, z_d)}$$
$$= \pi(x_{1:d-2}, z_{d-1}, z_d) \frac{\pi_{X_{d-1}|X^{-(d-1)}}(x_{d-1} \mid x_{1:d-2}, z_d)}{\pi_{X_{d-1}|X^{-(d-1)}}(z_{d-1} \mid x_{1:d-2}, z_d)}$$

hence

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

Proof.

By $z \in \text{supp}(\pi)$ we have that $\pi_{X_i)}(z_i) > 0$ for all i. Also, we are allowed to suppose that $\pi_{X_i}(x_i) > 0$ for all i. Thus all the conditional probabilities we introduce are positive since

$$\pi_{X_j|X^{-j}}(x_j \mid x_1, \dots, x_{j-1}, z_{j+1}, \dots, z_d)$$
$$= \frac{\pi(x_1, \dots, x_{j-1}, x_j, z_{j+1}, \dots, z_d)}{\pi(x_1, \dots, x_{j-1}, z_j, z_{j+1}, \dots, z_d)} > 0.$$

By iterating we have the theorem. □

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 $\iint \exp(-x_1 x_2)\, dx_1 dx_2 = \infty$ 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

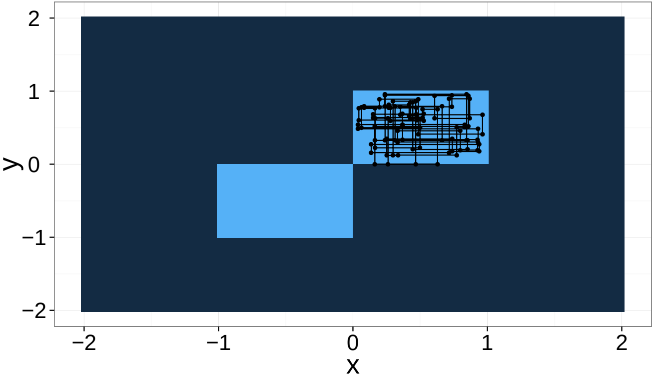


Figure: Gibbs sampling targeting
$\pi(x, y) \propto \mathbf{1}_{[-1,0]\times[-1,0]\cup[0,1]\times[0,1]}(x, y)$.

Invariance of the Gibbs sampler I

- 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.

Invariance of the Gibbs sampler II

Proof for $d = 2$.

We have

$$\int K(x,y)\pi(x)dx = \int \pi(y_2 \mid y_1)\pi(y_1 \mid x_2)\pi(x_1,x_2)dx_1dx_2$$
$$= \pi(y_2 \mid y_1) \int \pi(y_1 \mid x_2)\pi(x_2)dx_2$$
$$= \pi(y_2 \mid y_1)\pi(y_1) = \pi(y_1,y_2) = \pi(y).$$

□

Irreducibility and Recurrence

Proposition

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

Proof.

Irreducibility. Let $\mathbb{X} \subset \mathbb{R}^d$, such that $\pi(\mathbb{X}) = 1$. Write K for the kernel and let $A \subset \mathbb{X}$ such that $\pi(A) > 0$. Then for any $x \in \mathbb{X}$

$$\begin{aligned}
K(x, A) &= \int_A K(x, y)\mathrm{d}y \\
&= \int_A \pi_{X_1 | X^{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \\
&\qquad \times \pi_{X_d | X^{-d}}(y_d \mid y_1, \ldots, y_{d-1})\mathrm{d}y.
\end{aligned}$$

Proof.

Thus if for some $x \in \mathbb{X}$ and A with $\pi(A) > 0$ we have $K(x, A) = 0$, we must have that

$$\pi_{X_1|X^{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \times \pi_{X_d|X^{-d}}(y_d \mid x_1^{(t)}, \ldots, x_d^{(t)}) = 0,$$

for π-almost all $y = (y_1, \ldots, y_d) \in A$.

Therefore we must also have that

$$\pi\left(y_1, x_2, ..., y_d\right) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}\left(y_j \middle| y_{1:j-1}, x_{j+1:d}\right)}{\pi_{X_j|X_{-j}}\left(x_j \middle| y_{1:j-1}, x_{j+1:d}\right)} = 0,$$

for almost all $y = (y_1, \ldots, y_d) \in A$ and thus $\pi(A) = 0$ obtaining a contradiction.

Proof.

Recurrence. Recurrence follows from irreducibility and the fact that π is invariant. $\square$

CLT for Gibbs Sampler

Theorem

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

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

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

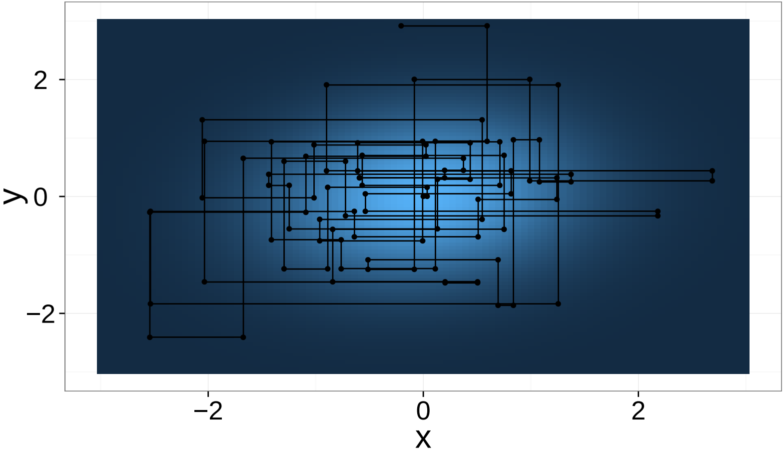


Figure: Case where $\rho = 0.1$, first 100 steps.

Bivariate Normal Distribution

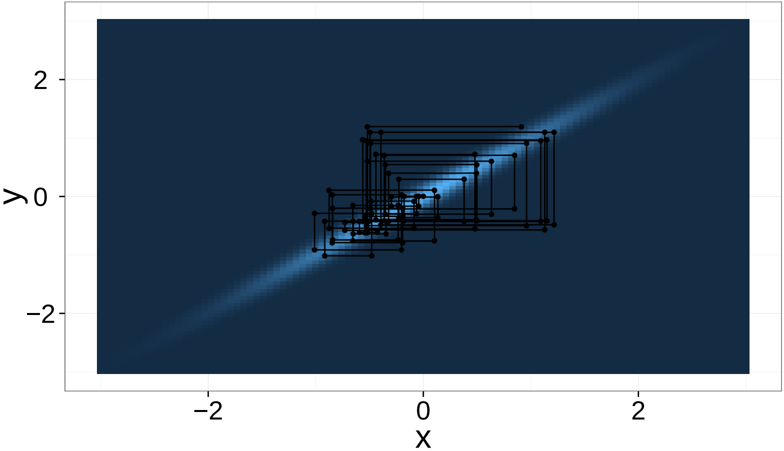
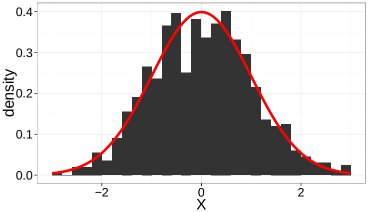
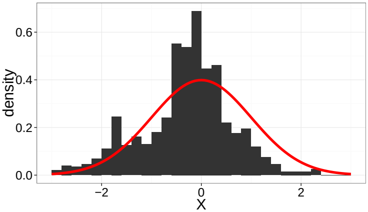


Figure: Case where $\rho = 0.99$, first 100 steps.

Bivariate Normal Distribution



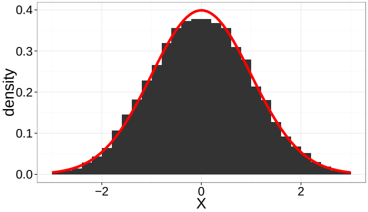
(a) Figure A



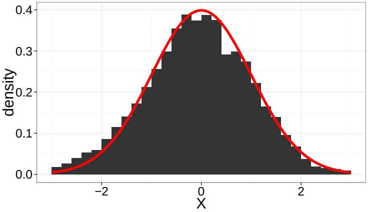
(b) Figure B

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



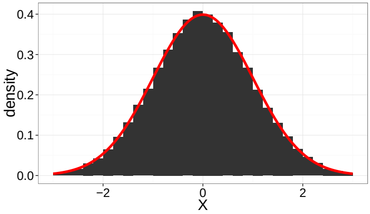
(a) b



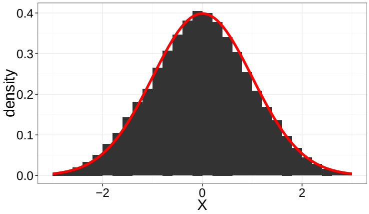
(b) b

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



(a) Figure A



(b) Figure B

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

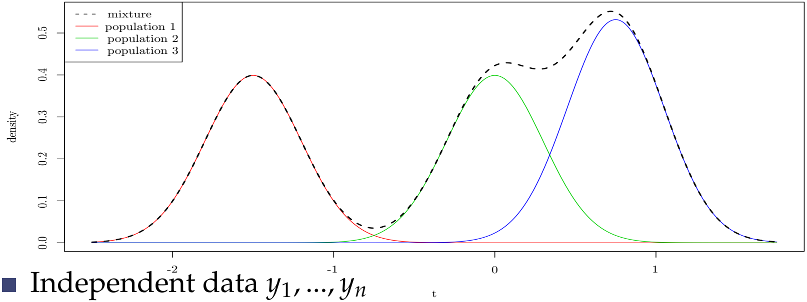
Gibbs Sampling and Auxiliary Variables

- 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 $\overline{\pi}\left(x_1, ..., x_d, z_1, ..., z_p\right)$ such that

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

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

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

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

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

$$\mathbb{P}\left(z_{1:n}\middle|\, y_{1:n},\theta\right)=\prod_{i=1}^{n}\mathbb{P}\left(z_i\middle|\, y_i,\theta\right)$$

where

$$\mathbb{P}\left(z_i\middle|\, y_i,\theta\right)=\frac{\mathbb{P}\left(z_i\middle|\,\theta\right)p\left(y_i\middle|\, z_i,\theta\right)}{\sum_{k=1}^{K}\mathbb{P}\left(z_i=k\middle|\,\theta\right)p\left(y_i\middle|\, z_i=k,\theta\right)}$$

- Let $n_k=\sum_{i=1}^{n}\mathbf{1}_{\{k\}}\left(z_i\right), n_k\overline{y}_k=\sum_{i=1}^{n}y_i\mathbf{1}_{\{k\}}\left(z_i\right)$ 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

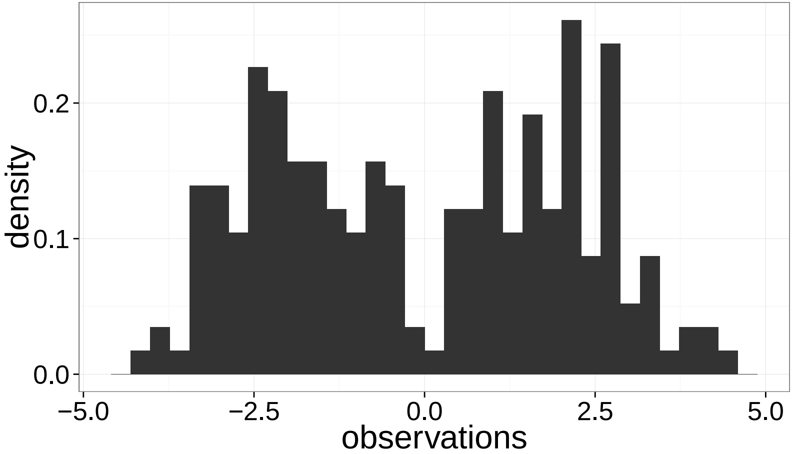


Figure: 200 points sampled from $\frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$.

Mixtures of Normals

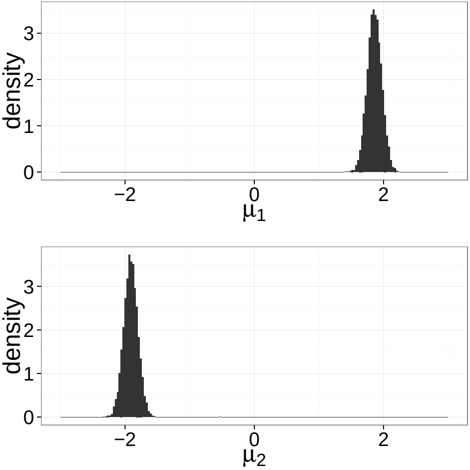


Figure: Histogram of the parameters obtained by 10,000 iterations of Gibbs sampling.

Mixtures of Normals

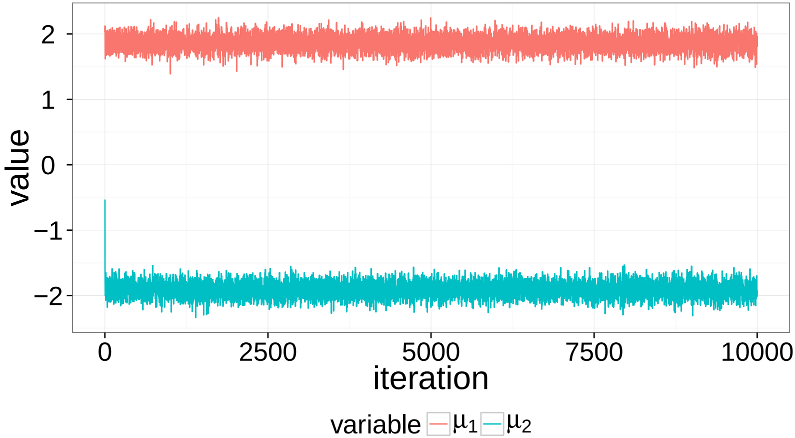


Figure: Traceplot of the parameters obtained by 10,000 iterations of Gibbs sampling.

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).