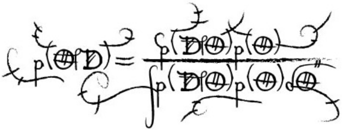


Bayesian Learning 732A46: Lecture 8

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

April 2016

Lecture overview

- Markov processes
- The concept of a stationary distribution of a Markov process
- The Gibbs sampler
- Data augmentation
  - Probit regression
  - Mixture models

Markov processes

- **For simplicity** consider a **discrete sample space** for θ. **Example**:

$$\pi(\theta) = \begin{cases} 1/4, & \text{if } \theta = \phi_1, \\ 7/12, & \text{if } \theta = \phi_2, \\ 1/6, & \text{if } \theta = \phi_3. \end{cases}$$

Definition

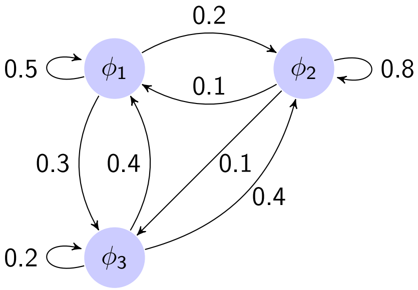
A *Markov* process is a collection of r.v's $\{\theta^{(t)}\}_{t\geq 0}$ with the property

$$\Pr(\theta^{(t)} = \phi^{(t)} | \theta^{(t-1)} = \phi^{(t-1)}, \ldots, \theta^{(1)} = \phi^{(1)}) = \Pr(\theta^{(t)} = \phi^{(t)} | \theta^{(t-1)} = \phi^{(t-1)}),$$

where $\phi^{(t)}$ denotes the state of the process at period t.
In the example with three states above: $\phi^{(t)} \in \{\phi_1, \phi_2, \phi_3\} \quad \forall t \geq 1$.

- A sequence generated by a Markov process is often called a **Markov chain**.
- **Discrete state space** gives us a thorough intuition. **Continuous state space** is a generalization.

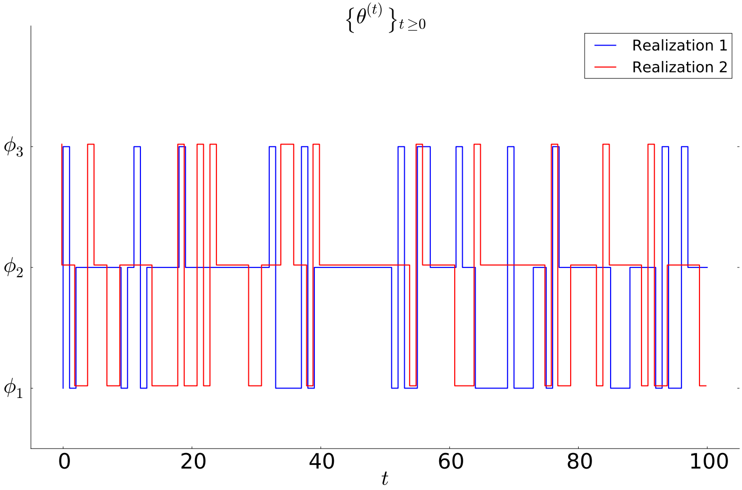
Transition probabilities



- **Transition probabilities** and **transition matrix**:

$$p_{ij} = \Pr(\theta^{(t)} = \phi_j | \theta^{(t-1)} = \phi_i) \quad \text{and} \quad P = \{p_{ij}\} = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.8 & 0.1 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$$

Simulating 100 draws from our process



Computing marginal distribution of the states at each t

- **Marginal distribution** at time t

$$\pi_j^{(t)} = \Pr(\theta^{(t)} = \phi_j).$$

- Let $\pi^{(0)} = (\pi_1^{(0)}, \pi_2^{(0)}, \pi_3^{(0)})$ denote the **initial state distribution**,

$$\pi_j^{(0)} = \Pr(\theta^{(0)} = \phi_j),$$

  i.e. the **marginal distribution** of state j at $t = 0$.

- What is the **marginal distribution** in $t = 1$ for state j?

$$\pi_j^{(1)} = \Pr(\theta^{(1)} = \phi_j) = \sum_i \underbrace{\Pr(\theta^{(1)} = \phi_j | \theta^{(0)} = \phi_i)}_{p_{ij}} \underbrace{\Pr(\theta^{(0)} = \phi_i)}_{\pi_i^{(0)}}.$$

- In **matrix form** $\pi^{(1)} = \pi^{(0)} P$.

- **What about $\pi_j^{(2)}$?**

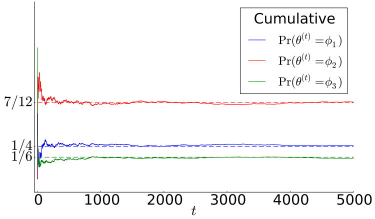
$$
\begin{aligned}
\pi_j^{(2)} &= \sum_i \sum_{i'} \Pr(\theta^{(2)} = \phi_j | \theta^{(1)} = \phi_i, \theta^{(0)} = \phi_{i'}) \Pr(\theta^{(1)} = \phi_i, \theta^{(0)} = \phi_{i'}) \\
&= \sum_i \sum_{i'} \underbrace{\Pr(\theta^{(2)} = \phi_j | \theta^{(1)} = \phi_i)}_{p_{ij}} \underbrace{\Pr(\theta^{(1)} = \phi_i | \theta^{(0)} = \phi_{i'})}_{p_{i'i}} \underbrace{\Pr(\theta^{(0)} = \phi_{i'})}_{\pi_{i'}^{(0)}} \\
&= \sum_i p_{ij} \underbrace{\sum_{i'} p_{i'i} \pi_{i'}^{(0)}}_{\pi_i^{(1)}}.
\end{aligned}
$$

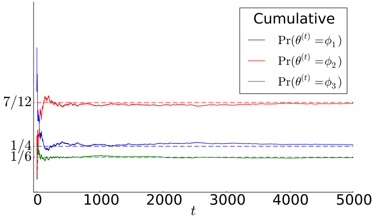
- **In fact**: $\pi^{(2)} = \pi^{(0)} P^2$ $(= \underbrace{\pi^{(0)} P}_{\pi^{(1)}} P = \pi^{(1)} P)$.
- **In general**: $\pi^{(n)} = \pi^{(0)} P^n$.

Stationary distribution of the states

- Suppose we observe the Markov process for an **infinite amount of time**.
  1. Does the **marginal distribution** of the states ever **stabilize**? In other words

$$\lim_{t\to\infty} \pi^{(t)} = \pi \quad \text{for some } \pi, \text{ regardless the initial } \pi^{(0)}?$$





  2. Is π **unique**?

Stationary distribution of the states, cont

- **In fact** (under conditions $(*)$, next slide),

$$\lim_{t \to \infty} P^t = \mathbb{1}\pi = \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}.$$

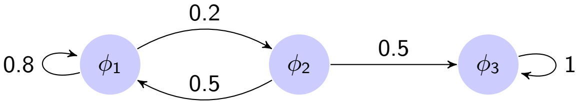
- **The stationary distribution** π, such that

$$\pi = \pi P,$$

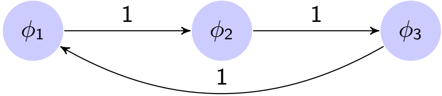
exists and is **unique** under $(*)$.

(∗) **The Markov chain** must be:

(i) **Irreducible**: **Positive probability** of reaching any state from any other state. **Not true for**:



(ii) **Aperiodic**: Does **not** get into "predictable cycles". **Predictable cycle**:



(iii) **Positive recurrent**: Expected time of returning to any state is **finite**. Define

$$T_i = \inf\{t \geq 1 : \theta^{(t)} = \phi_i | \theta^{(0)} = \phi_i\}.$$

The condition is

$$E[T_i] < \infty, \quad \text{for all states } i.$$

A sufficient condition to make life easier

- Conditions 1-3 are **necessary**. If true $\Rightarrow$ a unique stationary distribution.
- There exist a **"stronger property"** - **reversible Markov chain**. Important for **Metropolis-Hastings** (makes life easier!)

Definition

A **Markov chain** is **reversible** if there exist a distribution over the states, say π, such that

$$\Pr(\theta^{(t)} = \phi_j | \theta^{(t-1)} = \phi_i)\pi_i = \Pr(\theta^{(t)} = \phi_i | \theta^{(t-1)} = \phi_j)\pi_j, \tag{1}$$

for all t and all states i, j. Equation (1) is often called **the detailed balance condition**.

- For a **reversible chain**, π is always a **stationary distribution**:
  $\Pr(\theta^{(t)} = \phi_j) =$

$$\begin{aligned}\sum_i \Pr(\theta^{(t)} = \phi_j | \theta^{(t-1)} = \phi_i)\pi_i &\overset{(1)}{=} \sum_i \Pr(\theta^{(t)} = \phi_i | \theta^{(t-1)} = \phi_j)\pi_j \\ &= \pi_j \overbrace{\sum_i \Pr(\theta^{(t)} = \phi_i | \theta^{(t-1)} = \phi_j)}^{=1 \text{ (probability)}} = \pi_j\end{aligned}$$

- **Rewrite** Eq. (1) as

$$\begin{aligned}
\Pr(\theta^{(t)} = \phi_j | \theta^{(t-1)} = \phi_i)\pi_i &= \Pr(\theta^{(t)} = \phi_i | \theta^{(t-1)} = \phi_j)\pi_j \\
&\Leftrightarrow \\
\frac{\Pr(\theta^{(t)} = \phi_j, \theta^{(t-1)} = \phi_i)}{\Pr(\theta^{(t-1)} = \phi_i)}\pi_i &= \frac{\Pr(\theta^{(t)} = \phi_i, \theta^{(t-1)} = \phi_j)}{\Pr(\theta^{(t-1)} = \phi_j)}\pi_j
\end{aligned}$$

- **If we start the chain** at the stationary distribution π:

$$\Pr(\theta^{(t)} = \phi_i) = \pi_i \quad \text{and} \quad \Pr(\theta^{(t)} = \phi_j) = \pi_j$$

  for all t, thus $\Pr(\theta^{(t)} = \phi_j, \theta^{(t-1)} = \phi_i) = \Pr(\theta^{(t)} = \phi_i, \theta^{(t-1)} = \phi_j)$

- **In words:** The (unconditional) probability of going from $\phi_i \to \phi_j$ is the same as going from $\phi_j \to \phi_i$.
- **"Stronger property"**: There are Markov chains that are **not** reversible but still have a stationary distribution. **Reversibility is a sufficient** (but **not** necessary) condition.

Markov chains with continuous state space

- **Transition kernel** $T(\theta^{(t-1)} \to x)$ - a **conditional distribution** that expresses the probability to move to state x, **conditional** that the chain is at $\theta^{(t-1)}$.
- In **dicrete and finite state space** (what we have seen so far)

$$T_{ij}(\theta^{(t-1)} \to \theta^{(t)}) = \Pr(\theta^{(t)} = \phi_j | \theta^{(t-1)} = \phi_i)$$

- In **continuous state space**

$$T(\theta^{(t-1)} \to d\theta^{(t)}) = \Pr(d\theta^{(t)} | \theta^{(t-1)})$$

- $d\theta^{(t)} =$ Region in θ space.
- **Example (next lecture):** The **Metropolis-Hastings algorithm** uses the *detailed balance condition* to determine T. By construction this gives a chain that converges to $\pi(\theta) = p(\theta|y)$.

Computing expectations of a function using dependent draws

- **Recall:** we use the draws to estimate $I = E[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$ by

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)})$$

- The draws are **(Markov) dependent**... will it still work?
- **Yes**, in fact

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \xrightarrow{a.s} E[h(\theta)]$$

  still holds!
- The **statistical efficiency** of $\hat{I}$ is **reduced** because of the dependence.
- We have sacrificed the iid property and get **less efficient draws**. **In exchange**: we can handle **larger** dimensions of θ.

The Gibbs sampler

- Suppose the parameter vector is divided into K **blocks**

$$\theta = (\theta_1, \ldots, \theta_K).$$

- Each θ_k, $1 \leq k \leq K$ can be either a scalar or a vector itself.
- The Gibbs sampler is convenient when

$$\pi(\theta) = \pi(\theta_1, \ldots, \theta_K) \quad [= p(\theta_1, \ldots, \theta_K | y)]$$

  is **difficult to simulate**, but it is **easy to simulate** the **full conditional posteriors**

$$\pi(\theta_1 | \theta_2, \theta_3 \ldots, \theta_K)$$
$$\pi(\theta_2 | \theta_1, \theta_3 \ldots, \theta_K)$$
$$\vdots$$
$$\pi(\theta_K | \theta_1, \theta_2, \ldots, \theta_{K-1})$$

- **The Gibbs sampler** simulates from $\pi(\theta)$ by **alternating** the **full conditionals**.

The Gibbs sampler, cont.

The Gibbs sampler

Obtain N samples from $\pi(\theta)$.

- Set an (arbitrary) start point

$$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_K^{(0)}).$$

- **For** $i = 1, \ldots, N$, **repeat**

1. $\theta_1^{(i)} \sim \pi(\theta_1 | \theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_K^{(i-1)})$,
2. $\theta_2^{(i)} \sim \pi(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots \theta_K^{(i-1)})$,

$$\vdots$$

K. $\theta_K^{(i)} \sim \pi(\theta_K | \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{K-1}^{(i)})$.

- **Note**: in each draw, the **latest update** of each block is used.

Example: Simulating a bivariate Normal distribution

- **Note:** This examples is only for **illustration purposes**. There are **much more** efficient non-Markovian algorithms to do this.
- **Bivariate normal**

$$\theta = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2\end{bmatrix}\right)$$

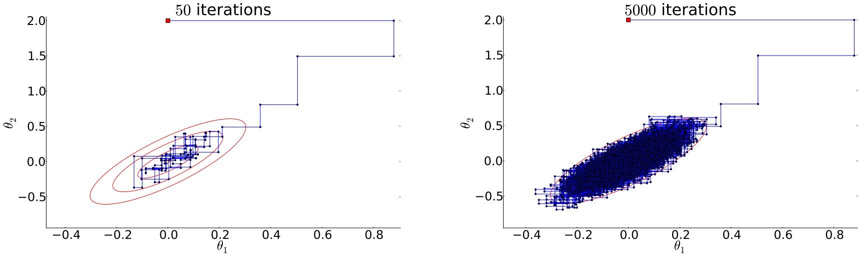
- **The full conditionals** (standard result for normal variates).

$$\begin{aligned}\theta_1|\theta_2 &\sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_2 - \mu_2), (1-\rho^2)\sigma_1^2\right) \\ \theta_2|\theta_1 &\sim \mathcal{N}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_1 - \mu_1), (1-\rho^2)\sigma_2^2\right).\end{aligned}$$

- Illustration (next slide) with

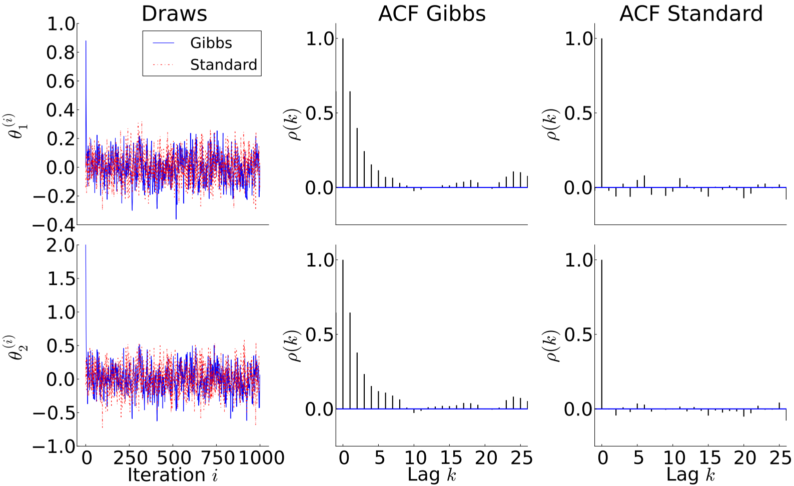
$$\mu_1 = \mu_2 = 0, \quad \sigma_1^2 = \sigma_2^2 = 1 \quad \text{and} \quad \rho = 0.5$$

- **Note:** The order of the full conditionals **does not matter**.



- The **contour plots** is the true $\pi(\theta_1, \theta_2)$.
- The **blue dots** are samples from $\pi(\theta_1, \theta_2)$ (after burn-in).
- Interested in the marginal of θ_1?. The **histogram** or **kernel density** of θ_1 approximates $\pi(\theta_1)$.
- The chain **"forgets"** the initial state.

- We compare to **direct** simulation, where θ_1, θ_2 are sampled jointly from a bivariate normal (and not dependent on previous draws). **More on measures of efficiency later**.



The power of Gibbs... and its drawback

- **Pros**:
  - Makes many hierarchical models a **piece of cake** to estimate.
  - **Data augmentation** - very powerful tool.
  - Appealing treatment of **missing data** problems.
- **Cons**:
  - **Inefficient if the blocks are correlated**. Takes a lot of time (**many draws**) to explore the posterior distribution.
- **Fighting** the **cons**:
  1. **Heavily correlated parameters** should always be included **in the same block**.
  2. A **re-parametrization** of the model can improve the efficiency.
  3. Introducing **extra parameters** in your model can **break** the correlation.

- **Notation**: $\theta_{\neg k} =$ all blocks except the kth.

- The **full conditional** of any block k is proportional to the **likelihood times the prior**:

$$\begin{aligned}\pi(\theta_k|\theta_{\neg k}) &= \frac{p(\theta, y)}{p(\theta_{\neg k}, y)} \\ &\propto p(\theta|y) \propto p(y|\theta)p(\theta),\end{aligned}$$

  where $\theta = (\theta_1, \ldots, \theta_K)$.

- **Strategy to derive the full conditional for θ_k**: throw away **everything that does not depend on θ_k** in $p(y|\theta)p(\theta)$. Choose a **conjugate prior** for the kth block (θ_k) if possible.

- Normal model with a **semi-conjugate** prior $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$,

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

- **The posterior** $\theta = (\mu, \sigma^2)$

$$\pi(\theta) \propto \left(\prod_{i=1}^{n} \mathcal{N}(y_i | \mu, \sigma^2) \right) \mathcal{N}(\mu | \mu_0, \tau_0^2) \text{Inv-}\chi^2(\sigma^2 | \nu_0, \sigma_0^2)$$

- **Full conditional posteriors**

$$\mu | \sigma^2, y \sim \mathcal{N}\left(\mu_n, \tau_n^2\right) \quad \text{[usual expressions for } \mu_n \text{ and } \tau_n^2\text{]}$$
$$\sigma^2 | \mu, y \sim \text{Inv-}\chi^2 \left(\nu_n, \frac{\nu_0 \sigma_0^2 + \sum_{i=1}^{n} (y_i - \mu)^2}{n + \nu_0} \right)$$

Gibbs sampling for AR processes

- **AR(p) process**

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

- Let $\phi = (\phi_1, ..., \phi_p)'$.
- **Prior**:
  - $\mu \sim$Normal
  - $\phi \sim$Multivariate Normal
  - $\sigma^2 \sim$Scaled Inverse χ^2.
- The **posterior** can be simulated by Gibbs sampling:
  - $\mu|\phi, \sigma^2, y \sim$ Normal
  - $\phi|\mu, \sigma^2, y \sim$ Multivariate Normal
  - $\sigma^2|\mu, \phi, y \sim$ Scaled Inverse χ^2

- A **Finite mixture** combines several densities to **flexibly** model data.
- The densities are called **components**.
- **Two-component mixture of normals** [MN(2)]

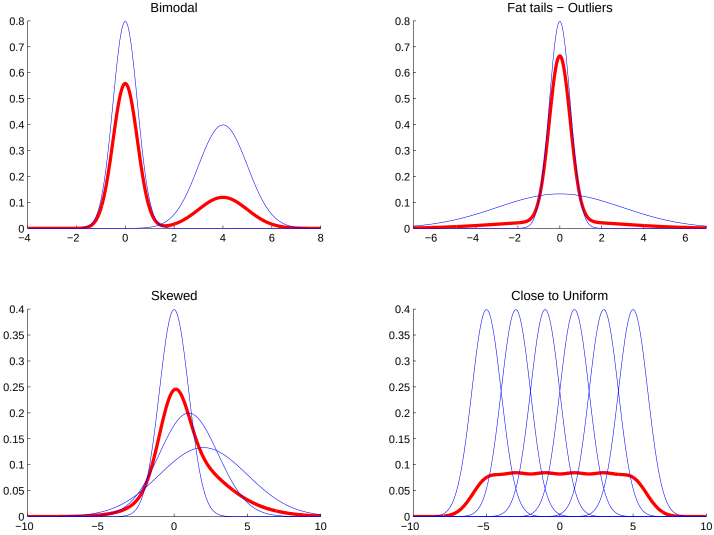
$$p(y|\mu, \sigma^2, \pi) = \pi_1 \cdot \mathcal{N}(y|\mu_1, \sigma_1^2) + \pi_2 \cdot \mathcal{N}(y|\mu_2, \sigma_2^2),$$

  where

$$\mu = (\mu_1, \mu_2), \sigma^2 = (\sigma_1^2, \sigma_2^2), \pi = (\pi_1, \pi_2) \quad \text{and } \pi_1 + \pi_2 = 1$$

- **Simulate** from a MN(2):
  1. Simulate an indicator $I \sim \mathrm{Bern}(\pi_1)$ with sample space $\{1, 2\}$.
  2. If $I = 1$, simulate y from $\mathcal{N}(\mu_1, \sigma_1^2)$ [$\pi_1 = \Pr(I = 1)$]
     If $I = 2$, simulate y from $\mathcal{N}(\mu_2, \sigma_2^2)$ [$\pi_2 = \Pr(I = 2)$].

Illustration of finite mixture distributions



Finite Mixture of normals

- **Not easy** to estimate **directly** - the likelihood is a product of sums.
- **Alternative** formulation of the model using the indicators I_i

$$\begin{aligned} \Pr(I_i = m|\pi_m) &= \pi_m \\ y_i|\mu_m, \sigma_m^2, I_i = m &\sim \mathcal{N}(\mu_m, \sigma_m^2). \end{aligned}$$

- **Assume** that we knew which of the two densities each observation came from.

$$I_i = \begin{cases} 1 \text{ if } y_i \text{ came from Density 1} \\ 2 \text{ if } y_i \text{ came from Density 2.} \end{cases}$$

- **Armed with knowledge** of $I_1, ..., I_n$ it is now **easy** to estimate π, $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ by **separating the sample** according to the I's.
- But we do **not** know $I_1, ..., I_n$!

Finite Mixture of normals, cont.

- **Gibbs sampling** to the rescue!
- **Assume:**
  1. Conjugate prior for $\pi \sim \mathrm{Beta}(\alpha_1, \alpha_2)$
  2. Conjugate prior for (μ_j, σ_j^2), see Lecture 5.
- Let $n_m = \sum_{i=1}^{n}(I_i == m)$, $m = 1, 2$, where

$$(I_i == m) = \begin{cases} 1 & \text{if } I_i = m \\ 0 & \text{if } I_i \neq m \end{cases} \quad \text{and } n_1 + n_2 = n.$$

- **Algorithm**:
  - $\pi \mid I, y \sim \mathrm{Beta}(\alpha_1 + n_1, \alpha_2 + n_2)$
  - $\sigma_1^2 \mid I, y \sim \text{Inv-}\chi^2$ and $\mu_1 | I, \sigma_1^2, y \sim \mathcal{N}$
  - $\sigma_2^2 \mid I, y \sim \text{Inv-}\chi^2$ and $\mu_2 | I, \sigma_2^2, y \sim \mathcal{N}$
  - $I_i \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, y \sim \mathrm{Bern}(\theta_i)$, $i = 1, ..., n$,

$$\theta_i = \frac{\pi_1 \mathcal{N}(y_i | \mu_1, \sigma_1^2)}{\pi_1 \mathcal{N}(y_i | \mu_1, \sigma_1^2) + (1 - \pi_1)\mathcal{N}(y_i | \mu_2, \sigma_2^2)}.$$

Finite Mixture of normals, cont.

- **Generalization**: K-component mixture of normals

$$p(y|\mu, \sigma^2, \pi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y|\mu_k, \sigma_k^2), \quad \text{where } \sum_{k=1}^{K} \pi_k = 1$$

- **Multi-class indicators**: $I_i = k$ if observation i comes from density k.
- **Gibbs sampling** (**Note** π: $\text{Beta} \to \text{Dirichlet}$, I_i: $\text{Bern} \to \text{Multinomial}$)
  - $(\pi_1, ..., \pi_K) \mid I, y \sim \text{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_K + n_K)$
  - $\sigma_k^2 \mid I, y \sim Inv\text{-}\chi^2$ and $\mu_k | I, \sigma_k^2, y \sim \mathcal{N}$, for $k = 1, ..., K$,
  - $I_i \mid \pi, \mu, \sigma^2, y \sim \text{Multinomial}(1; \theta_{i1}, ..., \theta_{iK})$, *for* $i = 1, ..., n$,

$$\theta_{ij} = \frac{\pi_j \mathcal{N}(y_i | mu_j, \sigma_j^2)}{\sum_{r=1}^{k} \pi_r \mathcal{N}(y_i | \mu_r, \sigma_r^2)}.$$

- We have **augmented** the model **with artificial data** $I = (I_1, \ldots, I_n)$.
- **Data augmentation**. **Downside**: increases the autocorrelation of the chain.

- **Probit** model:

$$\Pr(y_i = 1 \mid x_i) = \Phi(x_i'\beta) \quad [\Phi - \text{standard normal cdf}].$$

- **Random utility formulation** of the probit:

$$\begin{aligned} u_i &\sim \mathcal{N}(x_i'\beta, 1) \\ y_i &= \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases}. \end{aligned}$$

- This is an **equivalent** formulation:

$$\begin{aligned} \Pr(y_i = 1|x_i) = \Pr(u_i > 0) = 1 - \Pr(u_i \leq 0) = 1 - \Pr(u_i - x_i'\beta < -x_i'\beta) \\ = 1 - \Phi(-x_i'\beta) = \Phi(x_i'\beta). \end{aligned}$$

- If $u = (u_1, ..., u_n)$ were observed, then β could be analyzed by **standard linear regression** [**response**: u_i, **linear predictor** $x_i'\beta$, $\sigma^2 = 1$].
- But u is **not observed**... **Gibbs sampling** to the rescue!

Gibbs sampling for Probit regression

- Simulate from the **joint posterior** $p(u, \beta|y)$ alternating between the **full conditional posteriors**:
  1. $p(\beta|u, y)$, which is **multivariate normal** (just a **linear regression**)
  2. $p(u_i|\beta, y)$, $i = 1, ..., n$.
- The **full conditional posterior** distribution of u_i is:

$$
\begin{aligned}
p(u_i|\beta, y) &\propto p(y_i|\beta, u_i)p(u_i|\beta) \\
&= \begin{cases} \mathcal{N}(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (-\infty, 0] \text{ if } y_i = 0 \\ \mathcal{N}(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (0, \infty) \text{ if } y_i = 1. \end{cases}
\end{aligned}
$$

- Collect the β-draws. A **histogram** or **kernel density estimation** of these draws approximates

$$
p(\beta|y) = \int p(u, \beta|y) du.
$$