

Bayesian Learning

Lecture 7 - Gibbs sampling

Mattias Villani

Department of Statistics
Stockholm University



Lecture overview

- Monte Carlo simulation
- Gibbs sampling
- Data augmentation
 - ▶ Mixture models
 - ▶ Probit regression
- Regularized regression

Monte Carlo sampling

- If $\theta^{(1)}, \dots, \theta^{(m)}$ is an **iid sequence** from $p(\theta|\mathbf{y})$, then

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^m \theta^{(i)} \rightarrow \mathbb{E}(\theta|\mathbf{y})$$

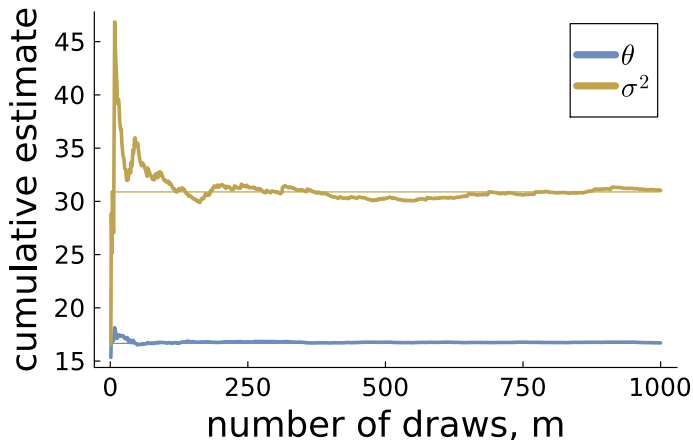
$$\bar{g}(\theta) = \frac{1}{m} \sum_{i=1}^m g(\theta^{(i)}) \rightarrow \mathbb{E}[g(\theta)|\mathbf{y}]$$

for some function $g(\theta)$ of interest.

- **Central limit theorem**

$$\bar{\theta}_{1:m} \stackrel{\text{appr}}{\sim} N\left(\mathbb{E}(\theta|\mathbf{y}), \frac{\mathbb{V}(\theta|\mathbf{y})}{m}\right) \quad \text{for large } m$$

Monte Carlo sampling - convergence



Gibbs sampling

- **Sampling from multivariate distributions**, $p(X_1, \dots, X_p)$.
- Typically a posterior distribution: $p(\theta_1, \dots, \theta_p | \mathbf{y})$.
- Requirement: Easily sampled **full conditional distributions**:
 - ▶ $p(\theta_1 | \theta_2, \theta_3, \dots, \theta_p, \mathbf{y})$
 - ▶ $p(\theta_2 | \theta_1, \theta_3, \dots, \theta_p, \mathbf{y})$
 - ▶ $\vdots$
 - ▶ $p(\theta_p | \theta_1, \theta_2, \dots, \theta_{p-1}, \mathbf{y})$
- Gibbs sampling is a special case of **Metropolis-Hastings**.
- Metropolis-Hastings is a **Markov Chain Monte Carlo (MCMC)** algorithm.

The Gibbs sampling algorithm

Gibbs sampling

Input: initial values $\theta_2^{(0)}, \dots, \theta_p^{(0)}$
number of posterior draws m .

for i in $1:m$ **do**

$$\theta_1 \sim p(\theta_1 \mid \theta_2^{(i-1)}, \theta_3^{(i-1)}, \dots, \theta_p^{(i-1)}, \mathbf{y})$$

$$\theta_2 \sim p(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i-1)}, \dots, \theta_p^{(i-1)}, \mathbf{y})$$

$$\vdots$$

$$\theta_p \sim p(\theta_p \mid \theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_{p-1}^{(i)}, \mathbf{y})$$

end

Output: m autocorrelated draws for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$
that converge in distribution to the joint
posterior $p(\theta_1, \dots, \theta_p \mid \mathbf{y})$.

Gibbs sampling draws converge to the posterior

- Gibbs draws $\theta^{(1)}, \dots, \theta^{(m)}$ are **dependent**, but

$$\bar{\theta} = \frac{1}{m} \sum_{t=1}^m \theta^{(t)} \rightarrow \mathbb{E}(\theta|\mathbf{y})$$

$$\bar{g}(\theta) = \frac{1}{m} \sum_{t=1}^m g(\theta^{(t)}) \rightarrow \mathbb{E}[g(\theta)|\mathbf{y}]$$

- $\theta^{(1)}, \dots, \theta^{(m)}$ **converges in distribution** to posterior $p(\theta|\mathbf{y})$.
- $\theta_j^{(1)}, \dots, \theta_j^{(m)}$ converges to the marginal posterior of θ_j .
- **Central limit theorem**

$$\bar{\theta} \overset{\text{approx}}{\sim} N(\mathbb{E}(\theta|\mathbf{y}), \mathbb{V}(\bar{\theta})) \text{ for large } m$$

Dependent draws for Gibbs are less efficient

- **Dependent draws** → **less efficient** than iid sampling.
- **IID samples:**

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{m}, \quad \text{where } \sigma^2 = \mathbb{V}(\theta|\mathbf{y})$$

- **Autocorrelated samples:**

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{m} \left(1 + 2 \sum_{k=1}^{\infty} \rho_k \right)$$

where ρ_k is the autocorrelation at lag k .

- **Inefficiency factor:**

$$\text{IF} = 1 + 2 \sum_{k=1}^{\infty} \rho_k \approx 1 + 2 \sum_{k=1}^K \rho_k$$

- **Effective sample size (ESS):** $\frac{m}{\text{IF}}$.

Gibbs sampling bivariate normal

■ Joint distribution

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

Gibbs sampling from a bivariate normal

Input: initial value $\theta_2^{(0)}$

number of posterior draws m .

for i in $1:m$ **do**

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

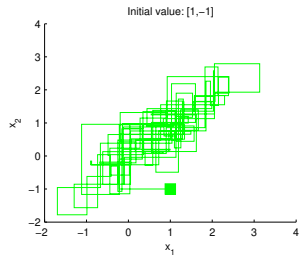
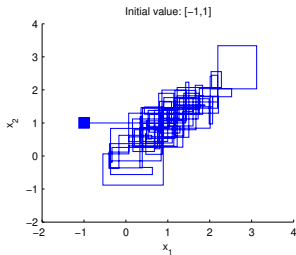
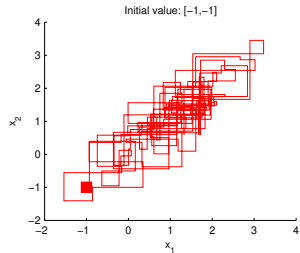
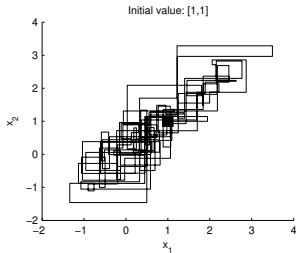
$$\left| \begin{array}{l} \theta_2^{(i)} \mid \theta_1 \sim N\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1} (\theta_1^{(i)} - \mu_1), \sigma_2^2 (1 - \rho)^2\right) \end{array} \right.$$

end

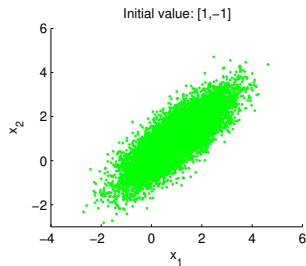
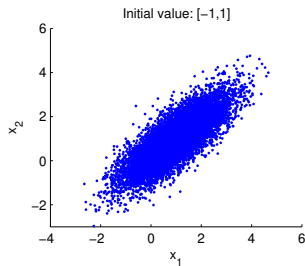
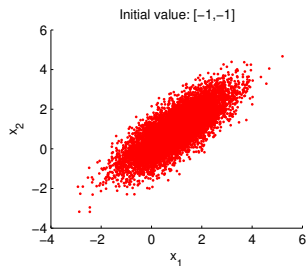
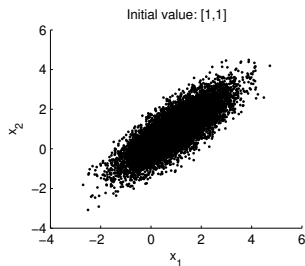
Output: m autocorrelated draws for $\theta = (\theta_1, \theta_2)^\top$ that converge in distribution to the bivariate normal distribution $\theta \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, \mu_2)^\top$ and

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

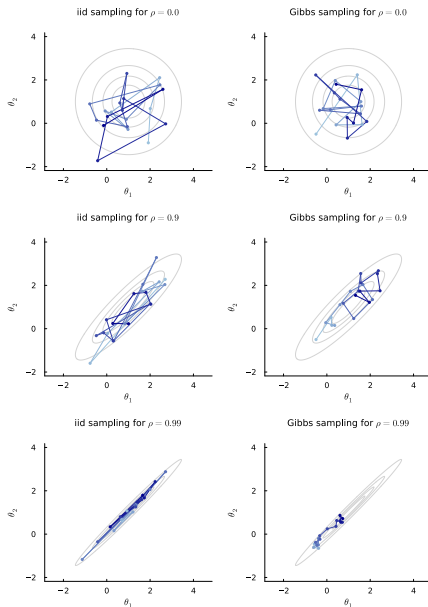
Gibbs sampling - Bivariate normal



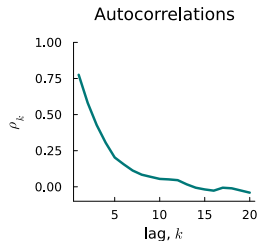
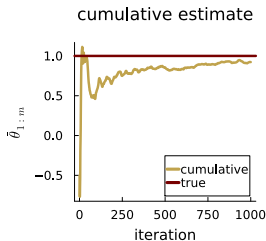
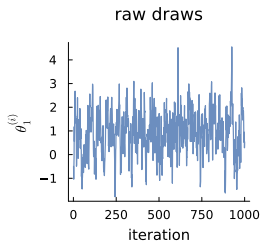
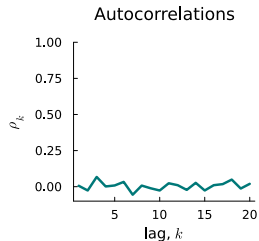
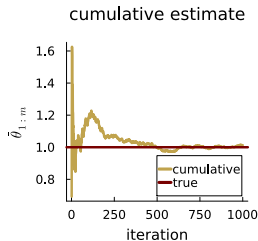
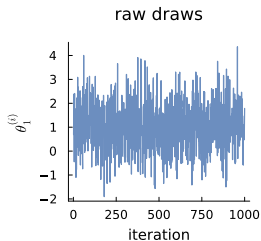
Gibbs sampling - Bivariate normal



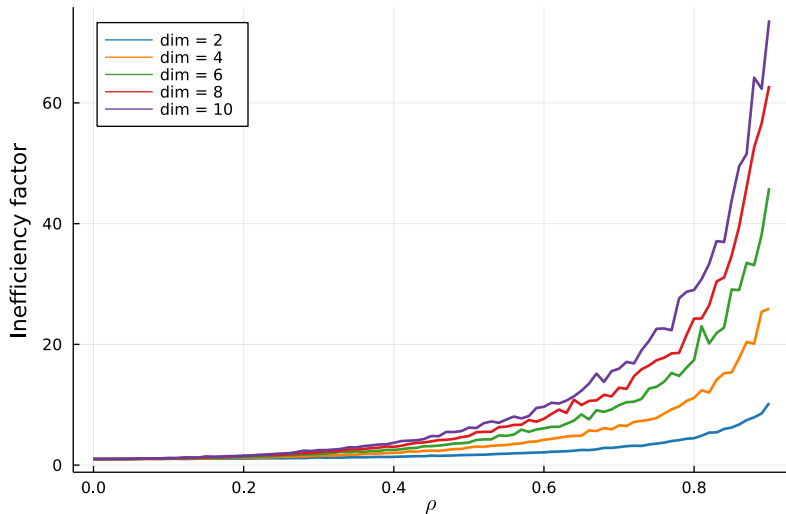
Direct sampling vs Gibbs sampling



Direct vs Gibbs sampling, bivariate normal $\rho = 0.9$



Gibbs is inefficient when parameters are correlated



Normal model with conditionally conjugate prior

■ Normal model with conditionally conjugate prior

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

■ Full conditional posteriors

$$\begin{aligned}\mu | \sigma^2, x &\sim N(\mu_n, \tau_n^2) \\ \sigma^2 | \mu, x &\sim \text{Inv} - \chi^2 \left(\nu_n, \frac{\nu_0 \sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{n + \nu_0} \right)\end{aligned}$$

with μ_n and τ_n^2 defined the same as when σ^2 is known.

Gibbs sampling for AR processes

■ AR(p) process

$$x_t = \mu + \phi_1(x_{t-1} - \mu) + \dots + \phi_p(x_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2).$$

■ Let $\phi = (\phi_1, \dots, \phi_p)'$.

■ Prior:

- ▶ $\mu \sim \text{Normal}$
- ▶ $\phi \sim \text{Multivariate Normal}$
- ▶ $\sigma^2 \sim \text{Scaled Inverse } \chi^2$.

■ The **posterior** can be simulated by **Gibbs sampling**:

- ▶ $\mu | \phi, \sigma^2, x \sim \text{Normal}$
- ▶ $\phi | \mu, \sigma^2, x \sim \text{Multivariate Normal}$
- ▶ $\sigma^2 | \mu, \phi, x \sim \text{Scaled Inverse } \chi^2$

Data augmentation - Mixture distributions

■ Let $N(x|\mu, \sigma^2)$ denote the **PDF** of $x \sim N(\mu, \sigma^2)$.

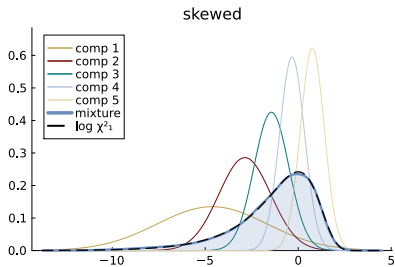
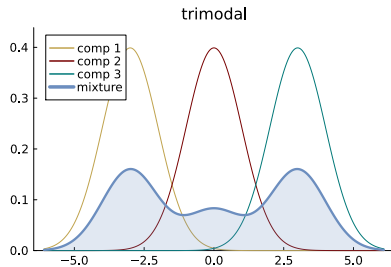
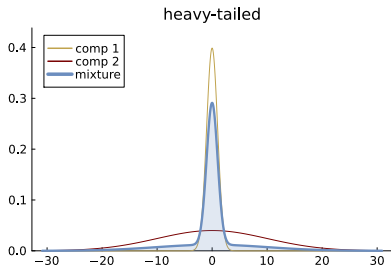
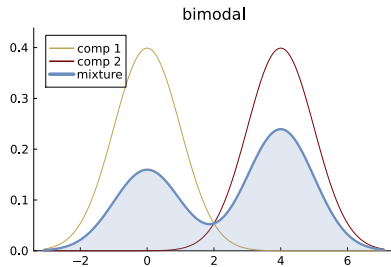
■ Two-component **mixture of normals** [MoN(2)]

$$p(x) = \omega \cdot N(x|\mu_1, \sigma_1^2) + (1 - \omega) \cdot N(x|\mu_2, \sigma_2^2)$$

■ **Simulate** from a MoN(2):

- ▶ Simulate a **membership indicator** $Z \in \{1, 2\}$: $Z \sim \text{Bern}(\pi)$.
- ▶ If $Z = 1$, simulate x from $N(\mu_1, \sigma_1^2)$
- ▶ If $Z = 2$, simulate x from $N(\mu_2, \sigma_2^2)$.

Illustration of mixture of normals



Data augmentation - Mixture distributions

■ K -component mixture of normals

$$p(x) = \sum_{k=1}^K \omega_k N(x|\mu_k, \sigma_k^2)$$

■ **Indicators:** $Z_i = k$ if observation x_i comes from component k .

Simulating data from a mixture of normals

Input: the number of simulated data observations n

mixture weights $\omega = (\omega_1, \dots, \omega_K)$

mixture component means $\mu_{1:K} = (\mu_1, \dots, \mu_K)$

mixture component variances $\sigma_{1:K}^2 = (\sigma_1^2, \dots, \sigma_K^2)$

for i in $1:n$ **do**

 // Simulate component allocation variable

 Draw $z_i \sim \text{Cat}(\omega_1, \dots, \omega_K)$

 // Simulate from selected mixture component

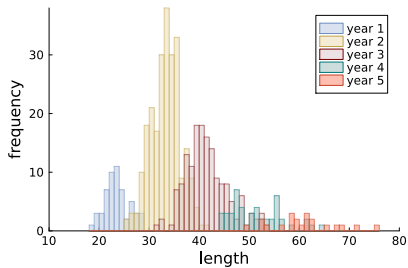
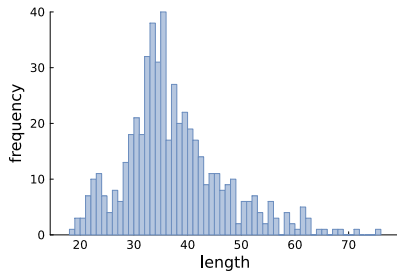
 Draw $x_i|z_i \sim N(\mu_{z_i}, \sigma_{z_i}^2)$

end

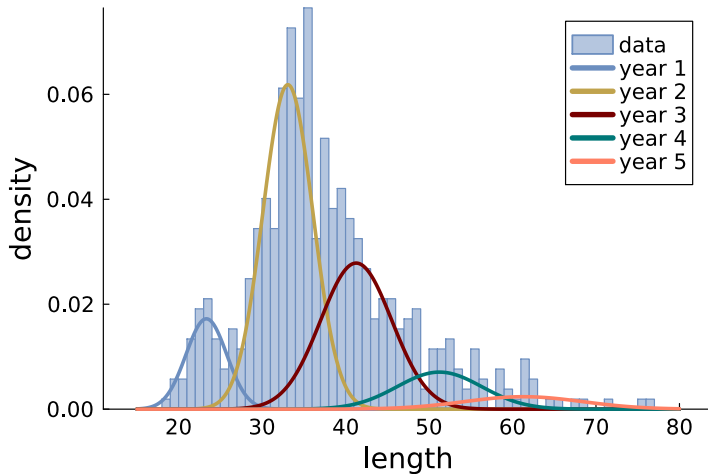
Output: n iid observations $\mathbf{x} = (x_1, \dots, x_n)$ from the mixture of normals model

$$p(x) = \sum_{k=1}^K \omega_k \cdot N(x|\mu_k, \sigma_k^2).$$

Fish length data with known yearly cohorts



Fish length data - fit with known yearly cohorts



Likelihood for a mixture and data augmentation

- The **likelihood** is a product of sums. **Messy** to work with.
- **Assume** that we know where each observation comes from

$z_i = k$ if x_i came from mixture component k .

- Given $z_1, \dots, z_n$ it is easy to estimate the means $\mu_1, \dots, \mu_K$, the variances $\sigma_1^2, \dots, \sigma_K^2$ and the mixture proportions $\omega_1, \dots, \omega_K$: just split up the data in K groups according to $z_1, \dots, z_n$.
- But we do **not** know $z_1, \dots, z_n$!
- **Data augmentation**: add $z_1, \dots, z_n$ as unknown parameters, and update them in separate Gibbs step.

Gibbs sampling for mixture distributions

```
for j in 1:m do
  // Update component parameters
  for k in 1:K do
    Set  $\mathbf{x}_k = \{x_i \text{ such that } z_i^{(j-1)} = k\}$ 
    Draw  $(\sigma_k^2)^{(j)} | \mathbf{x}_k \sim \text{ScaledInv-}\chi^2(\nu_{n,k}, \sigma_{n,k}^2)$ 
    Draw  $\mu_k^{(j)} | (\sigma_k^2)^{(j)}, \mathbf{x}_k \sim N(\mu_{n,k}, \tau_{n,k}^2)$ 
  end

  // Update component weights
  Set  $n_k = |\mathbf{x}_k|$ , number of obs in component k
  Draw  $\omega^{(j)} \sim \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_K + n_K)$ 

  // Update mixture allocations
  for i in 1:n do
    for k in 1:K do
       $\tilde{\omega}_k \propto \omega_k^{(j)} \cdot N(x_i | \mu_k^{(j)}, \sigma_k^{(j)})$ 
    end
    normalize  $\tilde{\omega}_1, \dots, \tilde{\omega}_K$  to sum to one
    simulate allocation  $z_i^{(j)} \sim \text{Cat}(\tilde{\omega}_1, \dots, \tilde{\omega}_K)$ 
  end
end
```

Fish length data - mixture of normals fit

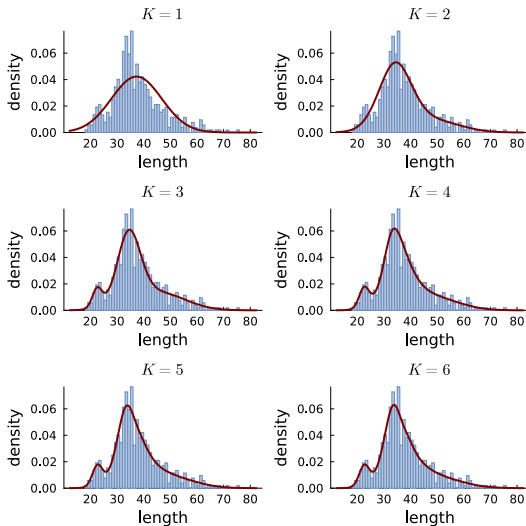
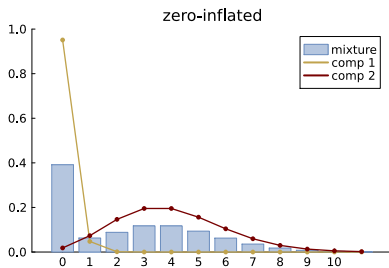
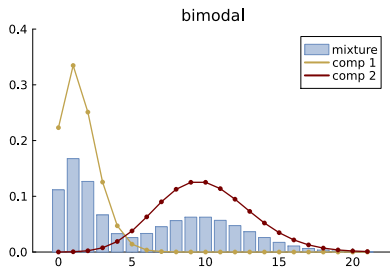
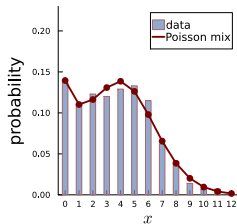
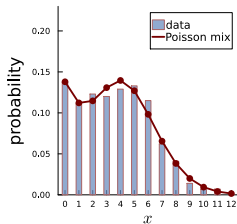
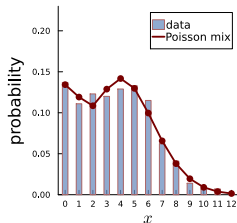
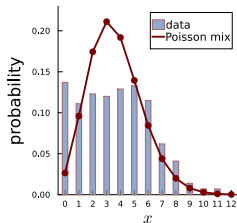


Illustration of mixture Poissons



Fitting a mixture Poissons to the eBay bidders



Data augmentation - Probit regression

■ Probit regression:

$$\Pr(y_i = 1 \mid \mathbf{x}_i) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$$

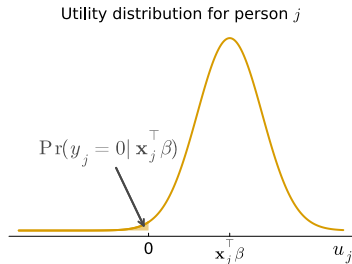
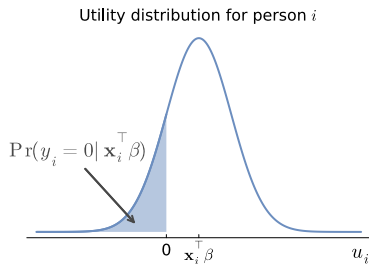
■ Random utility formulation:

$$\begin{aligned} u_i &\sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, 1) \\ y_i &= \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases} \end{aligned}$$

- Check: $\Pr(y_i = 1 \mid \mathbf{x}_i) = \Pr(u_i > 0) = 1 - \Pr(u_i \leq 0) = 1 - \Pr(u_i - \mathbf{x}_i^\top \boldsymbol{\beta} \leq -\mathbf{x}_i^\top \boldsymbol{\beta}) = 1 - \Phi(-\mathbf{x}_i^\top \boldsymbol{\beta}) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$.
- Given $\mathbf{u} = (u_1, \dots, u_n)$, $\boldsymbol{\beta}$ can be analyzed by linear regression.
- $\mathbf{u}$ is **not observed**. Gibbs sampling to the rescue!¹

¹Albert and Chib (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *JASA*.

Latent utility formulation of Probit regression



Gibbs sampling for the Probit regression

■ Random utility formulation:

$$u_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, 1)$$
$$y_i = \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases}$$

■ Simulate from **joint posterior** $p(\mathbf{u}, \boldsymbol{\beta} | \mathbf{y})$ by iterating between

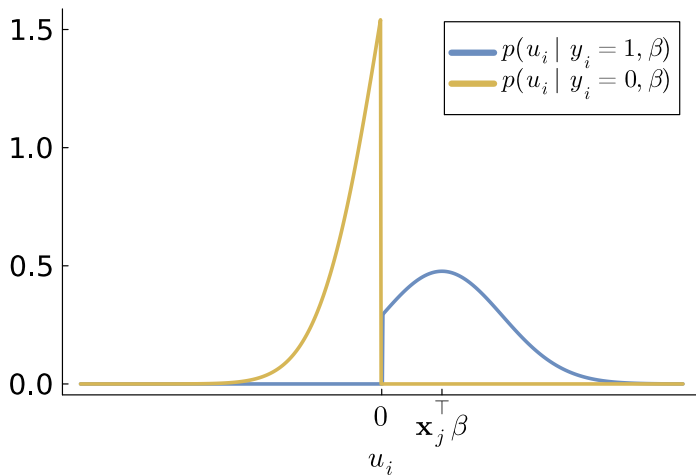
- ▶ $p(\boldsymbol{\beta} | \mathbf{u}, \mathbf{y})$ which is multivariate normal (linear regression)
- ▶ $p(u_i | \boldsymbol{\beta}, \mathbf{y})$, $i = 1, \dots, n$, independently.

■ The **full conditional** posterior distribution of u_i

$$p(u_i | \boldsymbol{\beta}, \mathbf{y}) \propto p(y_i | \boldsymbol{\beta}, u_i) p(u_i | \boldsymbol{\beta})$$
$$= \begin{cases} N(u_i | \mathbf{x}_i^\top \boldsymbol{\beta}, 1) & \text{truncated to } u_i \in (-\infty, 0] \text{ if } y_i = 0 \\ N(u_i | \mathbf{x}_i^\top \boldsymbol{\beta}, 1) & \text{truncated to } u_i \in (0, \infty) \text{ if } y_i = 1 \end{cases}$$

■ Histogram of $\boldsymbol{\beta}$ -draws approximates marginal posterior of $\boldsymbol{\beta}$.

Conditional posterior for latent utility u_i



Direct sampling L2-regularized regression

- Recap: The joint posterior of β , σ^2 and λ is

$$\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X} \sim N(\mu_n, \Omega_n^{-1})$$

$$\sigma^2|\lambda, \mathbf{y}, \mathbf{X} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$$

$$p(\lambda|\mathbf{y}, \mathbf{X}) \propto \sqrt{\frac{|\Omega_0|}{|\mathbf{X}'\mathbf{X} + \Omega_0|}} \left(\frac{\nu_n \sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)$$

- This is the **conditional-marginal decomposition**

$$p(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X}) = p(\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X})p(\sigma^2|\lambda, \mathbf{y}, \mathbf{X})p(\lambda|\mathbf{y}, \mathbf{X})$$

- **Gibbs sampling** can instead be used:

- ▶ Sample $\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X}$ from Normal
- ▶ Sample $\sigma^2|\beta, \lambda, \mathbf{y}, \mathbf{X}$ from $\text{Inv-}\chi^2$
- ▶ Sample $\lambda|\beta, \sigma^2, \mathbf{y}, \mathbf{X}$ from Gamma

- λ is **easy** to simulate **conditional on** β and σ^2 .

Gibbs sampling for L2-regularized regression

- Prior:

$$\begin{aligned}\beta|\sigma^2, \lambda &\sim N\left(\mathbf{0}, \frac{\sigma^2}{\lambda} I_k\right) \\ \sigma^2 &\sim \text{Inv} - \chi^2\left(\nu_0, \sigma_0^2\right) \\ \lambda^{-1} &\sim \text{Inv} - \chi^2\left(\omega_0, \psi_0^2\right).\end{aligned}$$

- By Bayes' theorem

$$p(\lambda|\beta, \sigma^2, \mathbf{y}) \propto p(\mathbf{y}|\beta, \sigma^2, \lambda) p(\lambda|\beta, \sigma^2)$$

- $p(\mathbf{y}|\beta, \sigma^2, \lambda)$ does not depend on λ once we condition on β :

$$p(\lambda|\beta, \sigma^2, \mathbf{y}) \propto p(\lambda|\beta, \sigma^2)$$

- So using Bayes' theorem once more

$$p(\lambda|\beta, \sigma^2, \mathbf{y}) \propto p(\lambda|\beta, \sigma^2) \propto p(\beta|\sigma^2, \lambda) p(\lambda)$$

- In conditional posterior for λ , the $\beta_1, \dots, \beta_p$ act like “data”.

Gibbs sampling for L2-regularized regression

Gibbs sampling linear regression - L2 regularization prior

The posterior for the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n), \quad (12.16)$$

with hierarchical L2 regularization prior

$$\begin{aligned}\boldsymbol{\beta} | \sigma^2, \psi^2 &\sim N(\mathbf{0}, \sigma^2 \psi^2 I_p) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \tau_0^2) \\ \psi^2 &\sim \text{Inv-}\chi^2(\omega_0, \psi_0^2).\end{aligned}$$

can be sampled by a two-block Gibbs sampler:

$$\begin{aligned}\text{Block1: } \boldsymbol{\beta} | \sigma^2, \psi^2, \mathbf{y} &\sim N(\hat{\boldsymbol{\beta}}_{L_2}, \sigma^2 (\mathbf{X}^\top \mathbf{X} + \psi^{-2} I_p)^{-1}) \\ \sigma^2 | \psi^2, \mathbf{y} &\sim \text{Inv-}\chi^2(\tau_n^2, \nu_n)\end{aligned}$$

$$\text{Block2: } \psi^2 | \boldsymbol{\beta}, \sigma^2, \mathbf{y} \sim \text{Inv-}\chi^2(\omega_n, \psi_n^2),$$

where $\hat{\boldsymbol{\beta}}_{L_2}$ is the ridge estimator

$$\hat{\boldsymbol{\beta}}_{L_2} = (\mathbf{X}^\top \mathbf{X} + \psi^{-2} I_p)^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^\top \mathbf{y}.$$

The hyperparameters ν_n and τ_n^2 are given in Figure 5.3.

Finally, $\omega_n = \omega_0 + p$ and $\psi_n^2 = (\sum_{i=1}^p (\beta_i / \sigma)^2 + \omega_0 \psi_0^2) / \omega_n$.

Improving the efficiency of the Gibbs sampler

- **Efficient blocking.** Correlated parameters should ideally be included in the same updating block.
- **Reparametrization.** Convergence can improve dramatically in alternative parametrizations.
- **Data augmentation.**
 - ▶ Augment with latent variables to make **full conditional posteriors more easily sampled** (Probit, Mixture models).
 - ▶ But typically **increases the autocorrelation** between draws.