Bayesian Learning

Lecture 7 - Gibbs sampling

Mattias Villani 👨

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Lecture overview

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

Monte Carlo sampling

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

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

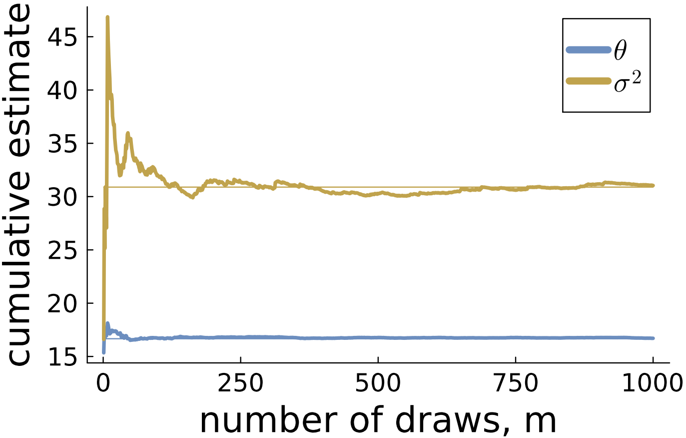
$$\bar{g}(\theta) = \frac{1}{m}\sum_{i=1}^{m}g(\theta^{(i)}) \quad \rightarrow \quad \mathbb{E}[g(\theta)|\boldsymbol{y}]$$

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

- **Central limit theorem**

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

Monte Carlo sampling - convergence



Gibbs sampling

- **Sampling from multivariate distributions**, $p(X_1, ..., X_p)$.
- Typically a posterior distribution: $p(\theta_1, \ldots, \theta_p | \boldsymbol{y})$.
- Requirement: Easily sampled **full conditional distributions**:
  - $p(\theta_1 | \theta_2, \theta_3 ..., \theta_p, \boldsymbol{y})$
  - $p(\theta_2 | \theta_1, \theta_3, ..., \theta_p, \boldsymbol{y})$
  - $\vdots$
  - $p(\theta_p | \theta_1, \theta_2, ..., \theta_{p-1}, \boldsymbol{y})$
- 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 $\theta_2^{(0)}, \ldots, \theta_p^{(0)}$
number of posterior draws m.

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

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

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

$$\vdots$$

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

end

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

Gibbs sampling draws converge to the posterior

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

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

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

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

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

Dependent draws for Gibbs are less efficient

- **Dependent draws** $\rightarrow$ **less efficient** than iid sampling.
- **IID samples**:

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

- **Autocorrelated samples**:

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

  where ρ_k is the autocorrelation at lag k.
- **Inefficiency factor**:

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

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

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**

$\theta_1^{(i)} \mid \theta_2 \sim \mathrm{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_2^{(i-1)} - \mu_2), \sigma_1^2(1-\rho)^2\right)$

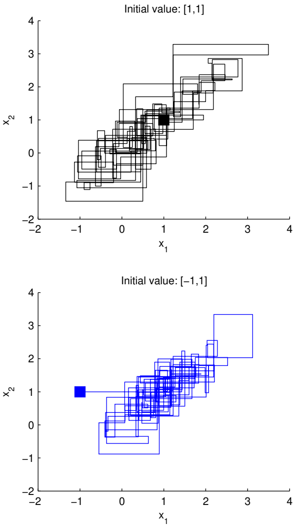
$\theta_2^{(i)} \mid \theta_1 \sim \mathrm{N}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_1^{(i)} - \mu_1), \sigma_2^2(1-\rho)^2\right)$

end

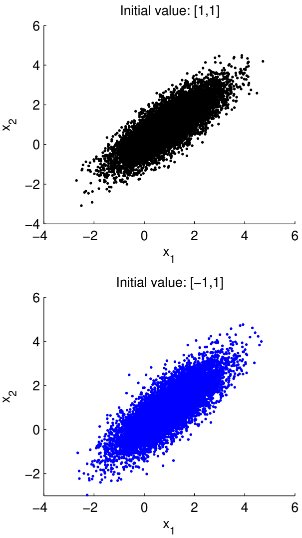
Output: m autocorrelated draws for $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$ that converge in distribution to the bivariate normal distribution $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and

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

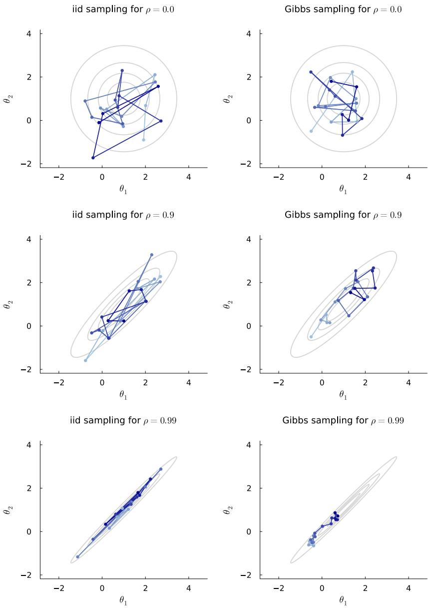
Gibbs sampling - Bivariate normal



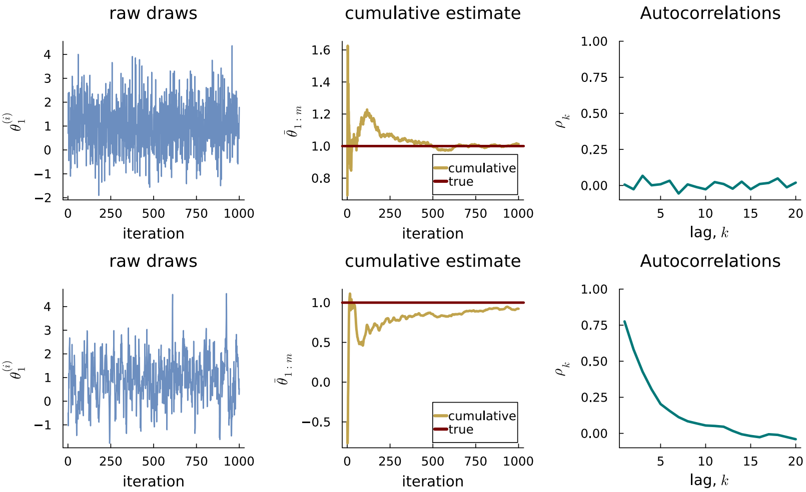
Gibbs sampling - Bivariate normal



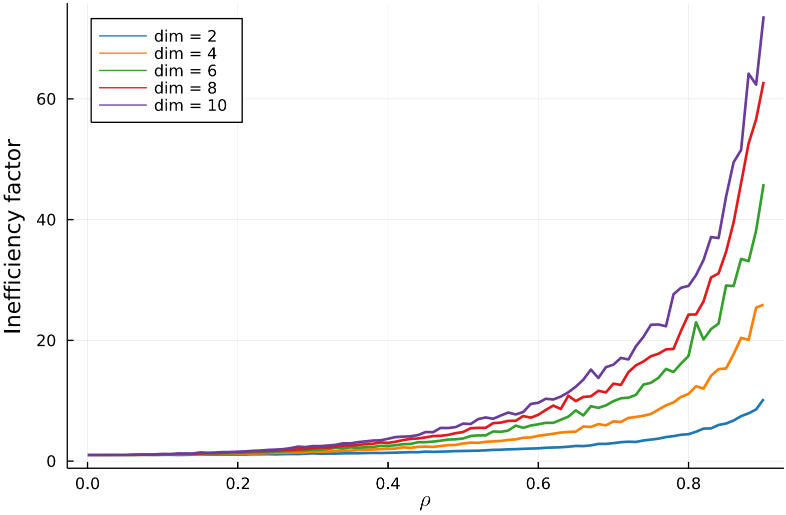
Direct sampling vs Gibbs sampling



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**

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

- **Full conditional posteriors**

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

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

Gibbs sampling for AR processes

- **AR(p) process**

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

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

Villani (2009). Steady State Priors for Vector Autoregressions. Journal of Applied Econometrics.

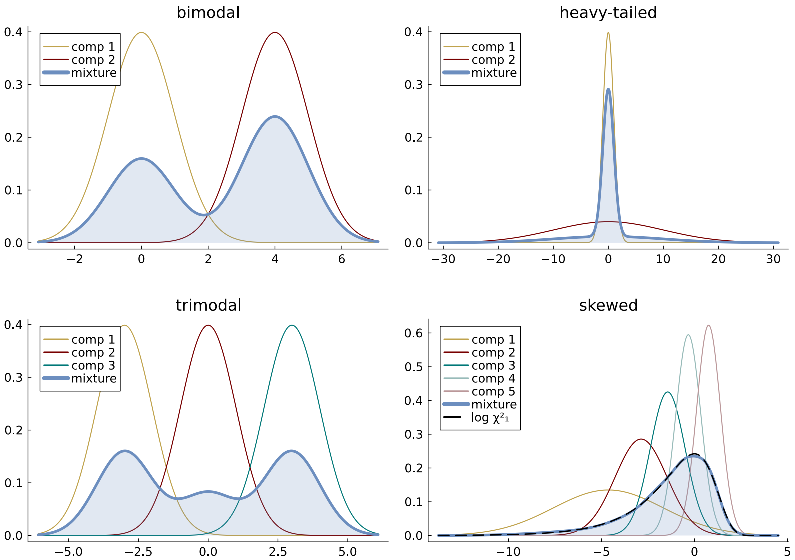
Data augmentation - Mixture distributions

- 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** $Z \in \{1, 2\}$: $Z \sim Bern(\pi)$.
  - 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 $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$
mixture component means $\boldsymbol{\mu}_{1:K} = (\mu_1, \ldots, \mu_K)$
mixture component variances $\boldsymbol{\sigma}^2_{1:K} = (\sigma_1^2, \ldots, \sigma_K^2)$

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

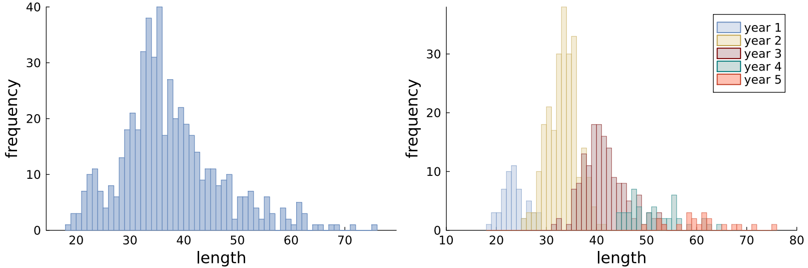
// Simulate component allocation variable
Draw $z_i \sim \text{Cat}(\omega_1, \ldots, \omega_K)$

// Simulate from selected mixture component
Draw $x_i|z_i \sim N(\mu_{z_i}, \sigma^2_{z_i})$

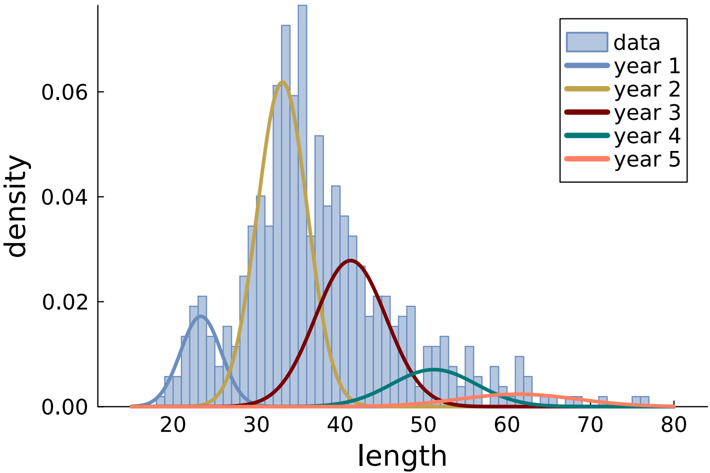
end

Output: n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ 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

$$z_i = k \text{ if } x_i \text{ came from mixture component } k.$$

- Given $z_1, ..., z_n$ it is easy to estimate the means $\mu_1, \ldots, \mu_K$, the variances $\sigma_1^2, \ldots, \sigma_K^2$ and the mixture proportions $\omega_1, \ldots \omega_K$: just split up the data in K groups according to $z_1, ..., z_n$.
- But we do **not** know $z_1, ..., z_n$!
- **Data augmentation**: add $z_1, ..., z_n$ 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 $\boldsymbol{\omega}^{(j)} \sim \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$

// 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 $\tilde{\omega}_1, \ldots, \tilde{\omega}_K$ to sum to one

simulate allocation $z_i^{(j)} \sim \text{Cat}(\tilde{\omega}_1, \ldots, \tilde{\omega}_K)$

end

Fish length data - mixture of normals fit

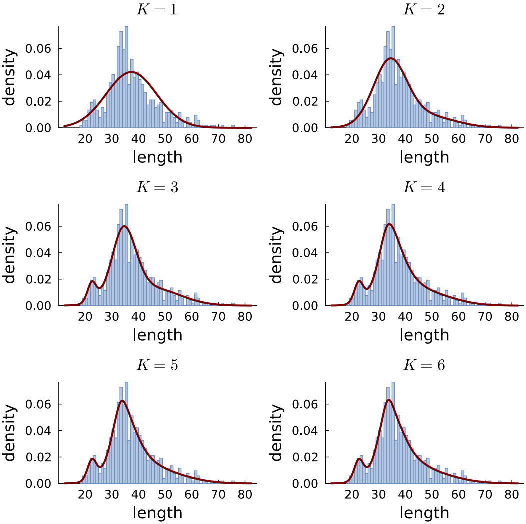
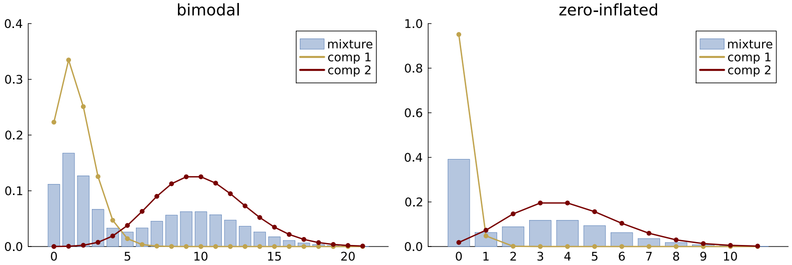
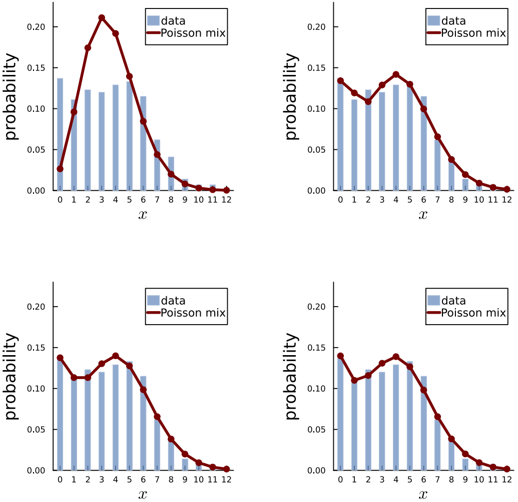


Illustration of mixture Poissons



Fitting a mixture Poissons to the eBay bidders



Data augmentation - Probit regression

- **Probit regression**:

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

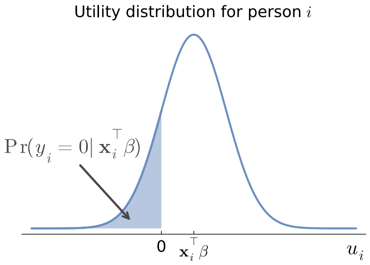
- **Random utility formulation**:

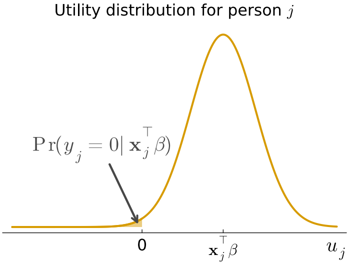
$$\begin{aligned} u_i &\sim N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1) \\ y_i &= \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases} \end{aligned}$$

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

[1] Albert 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**:

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

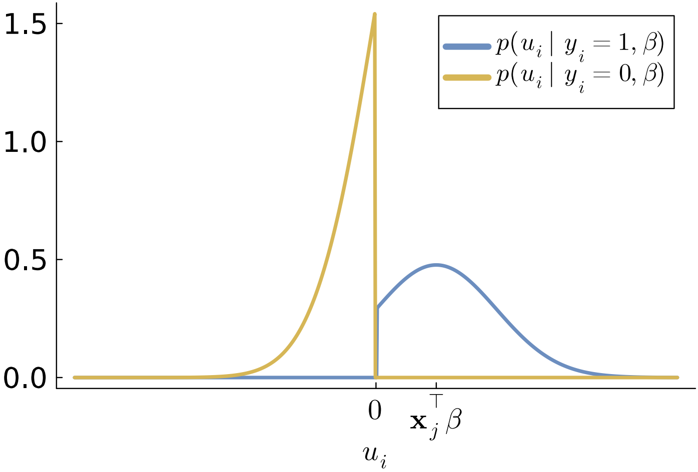
- Simulate from **joint posterior** $p(\boldsymbol{u}, \beta | \boldsymbol{y})$ by iterating between
  - $p(\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{y})$ which is multivariate normal (linear regression)
  - $p(u_i|\boldsymbol{\beta}, \boldsymbol{y})$, $i = 1, ..., n$, independently.

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

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

- Histogram of β-draws approximates marginal posterior of β.

Conditional posterior for latent utility u_i



Direct sampling L2-regularized regression

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

$$\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X} \sim N\left(\mu_n, \Omega_n^{-1}\right)$$
$$\sigma^2|\lambda, \mathbf{y}, \mathbf{X} \sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right)$$
$$p(\lambda|\mathbf{y}, \mathbf{X}) \propto \sqrt{\frac{|\Omega_0|}{|X'X + \Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)$$

- 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 Inv-χ^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:

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

- By Bayes' theorem

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

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

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

- So using Bayes' theorem once more

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

- In conditional posterior for λ, the $\beta_1, \ldots, \beta_p$ act like "data".

Gibbs sampling for L2-regularized regression

Gibbs sampling linear regression - L2 regularization prior

The posterior for the linear regression model

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

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:

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

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

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

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

The hyperparameters ν_n and τ_n^2 are given in Figure 5.3.
Finally, $\omega_n = \omega_0 + p$ and $\psi_n^2 = \left(\sum_{i=1}^{p}(\beta_i/\sigma)^2 + \omega_0\psi_0^2\right)/\omega_n$.

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.