Lecture 9: Gibbs and Data Augmentation

Merlise Clyde

September 29

Binary Regression

$$Y_i \mid \beta \sim \mathsf{Ber}(p(x_i^T \beta))$$

where $\Pr(Y_i = 1 \mid \beta) = p(x_i^T \beta))$ and linear predictor $x_i^T \beta = \lambda_i$

Binary Regression

$$Y_i \mid \beta \sim \text{Ber}(p(x_i^T \beta))$$

where $\Pr(Y_i = 1 \mid \beta) = p(x_i^T \beta))$ and linear predictor $x_i^T \beta = \lambda_i$

- link function for binary regression is any 1-1 function g that will map $(0, 1) \to \mathbb{R}$, i.e. $g(p(\lambda)) = \lambda$

Binary Regression

$$Y_i \mid \beta \sim \textsf{Ber}(p(x_i^T \beta))$$

where $\Pr(Y_i = 1 \mid \beta) = p(x_i^T \beta))$ and linear predictor $x_i^T \beta = \lambda_i$

- link function for binary regression is any 1-1 function g that will map $(0, 1) \to \mathbb{R}$, i.e. $g(p(\lambda)) = \lambda$
- logistic regression uses the logit link

$$\log\left(\frac{p(\lambda_i)}{1 - p(\lambda_i)}\right) = x_i^T \beta = \lambda_i$$

Binary Regression

$$Y_i \mid \beta \sim \text{Ber}(p(x_i^T \beta))$$

where $\Pr(Y_i = 1 \mid \beta) = p(x_i^T \beta))$ and linear predictor $x_i^T \beta = \lambda_i$

- link function for binary regression is any 1-1 function g that will map $(0, 1) \to \mathbb{R}$, i.e. $g(p(\lambda)) = \lambda$

- logistic regression uses the logit link

$$\log\left(\frac{p(\lambda_i)}{1 - p(\lambda_i)}\right) = x_i^T \beta = \lambda_i$$

- probit link

$$p(x_i^T \beta) = \Phi(x_i^T \beta)$$

- $\Phi()$ is the Normal cdf

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^{\mathcal{T}}\beta)^{y_i}(1 - \Phi(x_i^{\mathcal{T}}\beta))^{1-y_i}$$

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^{\mathcal{T}}\beta)^{y_i}(1 - \Phi(x_i^{\mathcal{T}}\beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^{\mathcal{T}} \beta)^{y_i} (1 - \Phi(x_i^{\mathcal{T}} \beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^T \beta)^{y_i} (1 - \Phi(x_i^T \beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$
- How to do approximate the posterior?

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^T \beta)^{y_i} (1 - \Phi(x_i^T \beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$
- How to do approximate the posterior?
  - asymptotic Normal approximation

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^{\mathcal{T}} \beta)^{y_i} (1 - \Phi(x_i^{\mathcal{T}} \beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$
- How to do approximate the posterior?
  - asymptotic Normal approximation
  - MH or adaptive Metropolis

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^T \beta)^{y_i} (1 - \Phi(x_i^T \beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$
- How to do approximate the posterior?
  - asymptotic Normal approximation
  - MH or adaptive Metropolis
  - stan (Hamiltonian Monte Carlo)

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^T \beta)^{y_i} (1 - \Phi(x_i^T \beta))^{1-y_i}$$

- prior $\beta \sim \mathrm{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$
- How to do approximate the posterior?
    - asymptotic Normal approximation
    - MH or adaptive Metropolis
    - stan (Hamiltonian Monte Carlo)
    - Gibbs ?

Likelihood and Posterior

Likelihood:

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} \Phi(x_i^{T}\beta)^{y_i}(1-\Phi(x_i^{T}\beta))^{1-y_i}$$

- prior $\beta \sim \mathsf{N}_p(b_0, \Phi_0)$
- posterior $\pi(\beta) \propto \pi(\beta)\mathcal{L}(\beta)$
- How to do approximate the posterior?
    - asymptotic Normal approximation
    - MH or adaptive Metropolis
    - stan (Hamiltonian Monte Carlo)
    - Gibbs ?

seemingly no, but there is a trick!

Data Augmentation

- Consider an **augmented** posterior

$$\pi(\beta, Z \mid y) \propto \pi(\beta)\pi(Z \mid \beta)\pi(y \mid Z, \theta)$$

Data Augmentation

- Consider an **augmented** posterior

$$\pi(\beta, Z \mid y) \propto \pi(\beta)\pi(Z \mid \beta)\pi(y \mid Z, \theta)$$

- IF we choose $\pi(Z \mid \beta)$ and $\pi(y \mid Z, \theta)$ carefully, we can carry out Gibbs and get samples of $\pi(\beta \mid y)$!

Data Augmentation

- Consider an **augmented** posterior

$$\pi(\beta, Z \mid y) \propto \pi(\beta)\pi(Z \mid \beta)\pi(y \mid Z, \theta)$$

- IF we choose $\pi(Z \mid \beta)$ and $\pi(y \mid Z, \theta)$ carefully, we can carry out Gibbs and get samples of $\pi(\beta \mid y)$!

$$\pi(\beta \mid y) = \int_{\mathcal{Z}} \pi(\beta, z \mid y)\, dz$$

(it is a marginal of joint augmented posterior)

Data Augmentation

- Consider an **augmented** posterior

$$\pi(\beta, Z \mid y) \propto \pi(\beta)\pi(Z \mid \beta)\pi(y \mid Z, \theta)$$

- IF we choose $\pi(Z \mid \beta)$ and $\pi(y \mid Z, \theta)$ carefully, we can carry out Gibbs and get samples of $\pi(\beta \mid y)$!

$$\pi(\beta \mid y) = \int_{\mathcal{Z}} \pi(\beta, z \mid y)\, dz$$

(it is a marginal of joint augmented posterior)

- We have to choose

$$p(y \mid \beta) = \int_{\mathcal{Z}} \pi(z \mid \beta)\pi(y \mid \beta, z)\, dz$$

Data Augmentation

- Consider an **augmented** posterior

$$\pi(\beta, Z \mid y) \propto \pi(\beta)\pi(Z \mid \beta)\pi(y \mid Z, \theta)$$

- IF we choose $\pi(Z \mid \beta)$ and $\pi(y \mid Z, \theta)$ carefully, we can carry out Gibbs and get samples of $\pi(\beta \mid y)$!

$$\pi(\beta \mid y) = \int_{\mathcal{Z}} \pi(\beta, z \mid y)\, dz$$

(it is a marginal of joint augmented posterior)

- We have to choose

$$p(y \mid \beta) = \int_{\mathcal{Z}} \pi(z \mid \beta)\pi(y \mid \beta, z)\, dz$$

- complete data likelihood

Augmentation Strategy

Set

- $y_i = 1_{(Z_i > 0)}$ i.e. ($y_i = 1$ if $Z_i > 0$)
- $y_i = 1_{(Z_i < 0)}$ i.e. ($y_i = 0$ if $Z_i < 0$)

Augmentation Strategy

Set

- $y_i = 1_{(Z_i > 0)}$ i.e. ($y_i = 1$ if $Z_i > 0$)
- $y_i = 1_{(Z_i < 0)}$ i.e. ($y_i = 0$ if $Z_i < 0$)
- $Z_i = x_i^T \beta + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim}$ N(0,1)

Augmentation Strategy

Set

- $y_i = 1_{(Z_i > 0)}$ i.e. ($y_i = 1$ if $Z_i > 0$)
- $y_i = 1_{(Z_i < 0)}$ i.e. ($y_i = 0$ if $Z_i < 0$)
- $Z_i = x_i^T \beta + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim}$ N(0,1)
- Relationship to probit model:

$$
\begin{aligned}
\Pr(y = 1 \mid \beta) &= P(Z_i > 0 \mid \beta) \\
&= P(Z_i - x_i^T \beta > -x^T \beta) \\
&= P(\epsilon_i > -x^T \beta) \\
&= 1 - \Phi(-x_i^T \beta) \\
&= \Phi(x_i^T \beta)
\end{aligned}
$$

Augmented Posterior & Gibbs

$$\pi(Z_1, \ldots, Z_n, \, \beta \mid y) \propto$$

$$\mathsf{N}(\beta; b_0, \phi_0) \left\{ \prod_{i=1}^{n} \mathsf{N}(Z_i; x_i^T \beta, 1) \right\} \left\{ \prod_{i=1}^{n} y_i 1_{(Z_i > 0)} + (1 - y_i) 1_{(Z_i < 0)} \right\}$$

Augmented Posterior & Gibbs

$$\pi(Z_1, \ldots, Z_n,\ \beta \mid y) \propto$$
$$\mathsf{N}(\beta; b_0, \phi_0) \left\{ \prod_{i=1}^{n} \mathsf{N}(Z_i; x_i^T \beta, 1) \right\} \left\{ \prod_{i=1}^{n} y_i 1_{(Z_i > 0)} + (1 - y_i) 1_{(Z_i < 0)} \right\}$$

- full conditional for β

$$\beta \mid Z_1, \ldots, Z_n, y_1, \ldots, y_n \sim \mathsf{N}(b_n, \Phi_n)$$

- standard Normal-Normal regression updating given Z_i's

Augmented Posterior & Gibbs

$$\pi(Z_1, \ldots, Z_n, \; \beta \mid y) \propto$$

$$\mathsf{N}(\beta; b_0, \phi_0) \left\{ \prod_{i=1}^{n} \mathsf{N}(Z_i; x_i^T \beta, 1) \right\} \left\{ \prod_{i=1}^{n} y_i 1_{(Z_i > 0)} + (1 - y_i) 1_{(Z_i < 0)} \right\}$$

- full conditional for β

$$\beta \mid Z_1, \ldots, Z_n, y_1, \ldots, y_n \sim \mathsf{N}(b_n, \Phi_n)$$

- standard Normal-Normal regression updating given Z_i's
- Full conditional for latent Z_i

$$\pi(Z_i \mid \beta, Z_{[-i]}, y_1, \ldots, y_n) \propto \mathsf{N}(Z_i; x_i^T \beta, 1) 1_{(Z_i > 0)} \text{ if } y_1 = 1$$
$$\pi(Z_i \mid \beta, Z_{[-i]}, y_1, \ldots, y_n) \propto \mathsf{N}(Z_i; x_i^T \beta, 1) 1_{(Z_i < 0)} \text{ if } y_1 = 0$$

- sample from independent truncated normal distributions !

Augmented Posterior & Gibbs

$$\pi(Z_1, \ldots, Z_n,\ \beta \mid y) \propto$$
$$\mathsf{N}(\beta; b_0, \phi_0) \left\{ \prod_{i=1}^n \mathsf{N}(Z_i; x_i^T \beta, 1) \right\} \left\{ \prod_{i=1}^n y_i 1_{(Z_i > 0)} + (1 - y_i) 1_{(Z_i < 0)} \right\}$$

- full conditional for β

$$\beta \mid Z_1, \ldots, Z_n, y_1, \ldots, y_n \sim \mathsf{N}(b_n, \Phi_n)$$

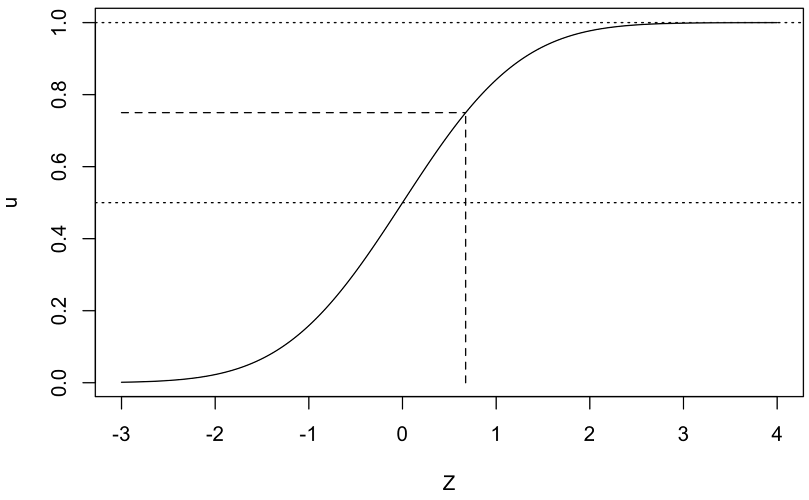
- standard Normal-Normal regression updating given Z_i's
- Full conditional for latent Z_i

$$\pi(Z_i \mid \beta, Z_{[-i]}, y_1, \ldots, y_n) \propto \mathsf{N}(Z_i; x_i^T \beta, 1) 1_{(Z_i > 0)} \text{ if } y_1 = 1$$
$$\pi(Z_i \mid \beta, Z_{[-i]}, y_1, \ldots, y_n) \propto \mathsf{N}(Z_i; x_i^T \beta, 1) 1_{(Z_i < 0)} \text{ if } y_1 = 0$$

- sample from independent truncated normal distributions !
- two block Gibbs sampler $\theta_{[1]} = \beta$ and $\theta_{[2]} = (Z_1, \ldots, Z_n)^T$

Truncated Normal Sampling

- Use inverse cdf method for cdf F
- If $u \sim U(0,1)$ set $z = F^{-1}(u)$



- if $Z \in (a,b)$, Draw $u \sim U(F(a), F(b))$ and set $z = F^{-1}(u)$

Data Augmentation in General

DA is a broader than a computational trick allowing Gibbs sampling

Data Augmentation in General

DA is a broader than a computational trick allowing Gibbs sampling

- missing data

Data Augmentation in General

DA is a broader than a computational trick allowing Gibbs sampling

- missing data
- random effects or latent variable modeling i.e we introduce latent variables to simplify dependence structure modelling

Data Augmentation in General

DA is a broader than a computational trick allowing Gibbs sampling

- missing data
- random effects or latent variable modeling i.e we introduce latent variables to simplify dependence structure modelling
- Modeling heavy tailed distributions such as t errors in regression

Comments

- Why don't we treat each individual θ_j as a separate block?

Comments

- Why don't we treat each individual θ_j as a separate block?
- Gibbs always accepts, but can mix slowly if parameters in different blocks are highly correlated!

Comments

- Why don't we treat each individual θ_j as a separate block?
- Gibbs always accepts, but can mix slowly if parameters in different blocks are highly correlated!
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)

Comments

- Why don't we treat each individual θ_j as a separate block?
- Gibbs always accepts, but can mix slowly if parameters in different blocks are highly correlated!
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)
- Collapse the sampler by integrating out as many parameters as possible (as long as resulting sampler has good mixing)

Comments

- Why don't we treat each individual θ_j as a separate block?
- Gibbs always accepts, but can mix slowly if parameters in different blocks are highly correlated!
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)
- Collapse the sampler by integrating out as many parameters as possible (as long as resulting sampler has good mixing)
- can use Gibbs steps and (adaptive) Metropolis Hastings steps together

Comments

- Why don't we treat each individual θ_j as a separate block?
- Gibbs always accepts, but can mix slowly if parameters in different blocks are highly correlated!
- Use block sizes in Gibbs that are as big as possible to improve mixing (proven faster convergence)
- Collapse the sampler by integrating out as many parameters as possible (as long as resulting sampler has good mixing)
- can use Gibbs steps and (adaptive) Metropolis Hastings steps together
- latent variables to allow Gibbs steps but not always better!