

Lecture 11: More Gibbs: Regression and Data Augmentation

Merlise Clyde

September 30



Normal Linear Regression Example

Model

$$\begin{aligned} Y \mid \beta, \phi &\sim \mathbf{N}(X\beta, \phi^{-1}I_n) \\ \beta &\sim \mathbf{N}(b_0, \Phi_0^{-1}) \\ \phi &\sim \text{Gamma}(v_0/2, s_0/2) \end{aligned}$$

- Full Conditional for β

$$\begin{aligned} \beta \mid \phi, y_1, \dots, y_n &\sim \mathbf{N}(b_n, \Phi_n^{-1}) \\ b_n &= (\Phi_0 + \phi X^T X)^{-1}(\Phi_0 b_0 + \phi X^T X \hat{\beta}) \\ \Phi_n &= \Phi_0 + \phi X^T X \end{aligned}$$

- Full Conditional for ϕ

$$\phi \mid \beta, y_1, \dots, y_n \sim \text{Gamma}((v_0 + n)/2, (s_0 + \sum_i (y_i - x_i^T \beta)))$$

- Choice of Prior Precision Φ_0



Invariance and Choice of Mean/Precision

- What if we transform the X matrix by $\tilde{X} = XH$ where H is $p \times p$ and invertible
- obtain the posterior for $\tilde{\beta}$ using Y and \tilde{X}

$$Y \sim N_n(\tilde{X}\tilde{\beta}, \phi^{-1}I_n)$$

- invariance suggests that since $\tilde{X}\tilde{\beta} = XH\tilde{\beta} = X\beta$ the posterior for β and $H\tilde{\beta}$ should be the same (or the posterior of $H^{-1}\beta$ and $\tilde{\beta}$ should be the same)
- with some linear algebra we can show that this is true if $b_0 = 0$ and Φ_0 is kX^TX for some k (show!)



Zellner's g-prior

Popular choice is to take $k = \phi/g$ which is a special case of Zellner's g-prior

$$\beta \mid \phi, g \sim N \left(0, \frac{g}{\phi} (X^T X)^{-1} \right)$$

- Full conditional

$$\beta \mid \phi, g \sim N \left(\frac{g}{1+g} \hat{\beta}, \frac{1}{\phi} \frac{g}{1+g} (X^T X)^{-1} \right)$$

- one parameter g controls shrinkage

if $\phi \sim \text{Gamma}(v_0/2, s_0/2)$ then posterior is

$$\phi \mid y_1, \dots, y_n \sim \text{Gamma}(v_n/2, s_n/2)$$

Conjugate so we could skip Gibbs sampling and sample directly from gamma and then conditional normal!



Ridge Regression

If $X^T X$ is nearly singular, certain elements of β or (linear combinations of β) may have huge variances under the g -prior (or flat prior) as the MLEs are highly unstable!

Ridge regression protects against the explosion of variances and ill-conditioning with the conjugate priors:

$$\beta \mid \phi \sim \mathbf{N}\left(0, \frac{1}{\phi\lambda} I_p\right)$$

Posterior for β (conjugate case)

$$\beta \mid \phi, \lambda, y_1, \dots, y_n \sim \mathbf{N}\left((\lambda I_p + X^T X)^{-1} X^T Y, \frac{1}{\phi} (\lambda I_p + X^T X)^{-1}\right)$$



Bayes Regression

- Posterior mean (or mode) given λ is biased, but can show that there **always** is a value of λ where the frequentist's expected squared error loss is smaller for the Ridge estimator than MLE!
- related to penalized maximum likelihood estimation
- Choice of λ
- usual center and standardized x !
- Bayes Regression and choice of Φ_0 in general is a very important problem and provides the foundation for many variations on shrinkage estimators, variable selection, hierarchical models, nonparameteric regression and more!
- Be sure that you can derive the full conditional posteriors for β and ϕ as well as the joint posterior in the conjugate case!



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 $(0, 1) \rightarrow \mathbb{R}$, i.e. $g(p(\lambda)) = \lambda$
- logistic regression use 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:

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

- prior $\beta \sim N_p(b_0, \Phi_0)$
- 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

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

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

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

(it is a marginal of joint augmented posterior)

- We have to choose

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

- 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 $Z_i < 0$)
- $Z_i = x_i^T \beta + \epsilon_i \quad \epsilon_i \stackrel{iid}{\sim} N(0,1)$
- Relationship to probit model:

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



Augmented Posterior & Gibbs

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

- full conditional for β

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

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

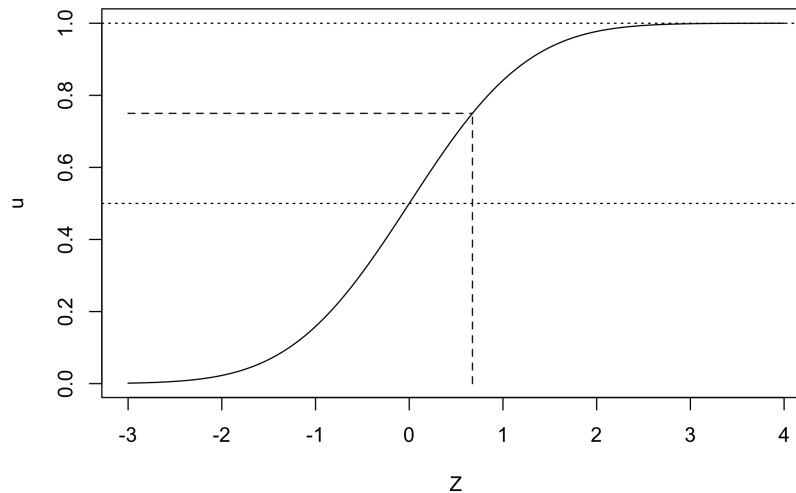
$$\begin{aligned} \pi(Z_i \mid \beta, Z_{[-i]}, y_1, \dots, y_n) &\propto \mathbf{N}(Z_i; x_i^T \beta, 1) 1(Z_i > 0) \text{ if } y_1 = 1 \\ \pi(Z_i \mid \beta, Z_{[-i]}, y_1, \dots, y_n) &\propto \mathbf{N}(Z_i; x_i^T \beta, 1) 1(Z_i < 0) \text{ if } y_1 = 0 \end{aligned}$$

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



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

- 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 β_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!

