Bayesian auxiliary variable models for binary and multinomial regression

(*Bayesian Analysis*, 2006)

Authors: Chris Holmes
Leonhard Held

As interpreted by: Rebecca Ferrell

UW Statistics 572, Talk #1

April 10, 2014

Categorical data setup

Classical framework with binary responses:

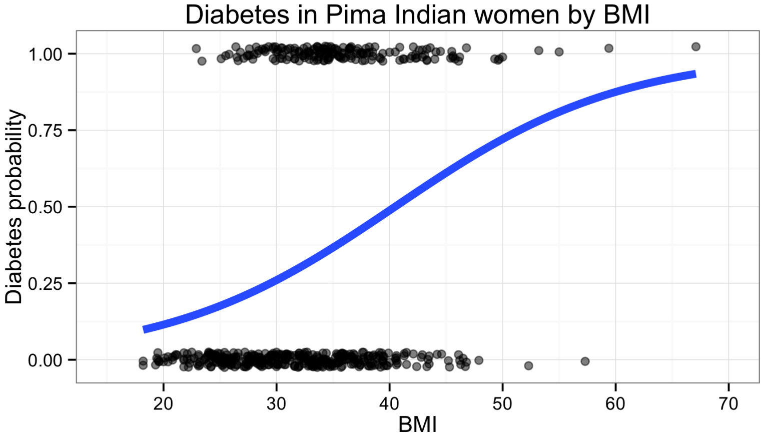
$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ p_i &= g^{-1}(\eta_i),\ g^{-1} : \mathbb{R} \to (0,1) \\ \eta_i &= \mathbf{x}_i \boldsymbol{\beta},\ i = 1, \ldots, n \\ \mathbf{x}_i &= \begin{pmatrix} x_{i1} & \ldots & x_{ip} \end{pmatrix} \\ \boldsymbol{\beta} &= \begin{pmatrix} \beta_1 & \ldots & \beta_p \end{pmatrix}^T \end{aligned}$$

Put a prior on the unknown coefficients:

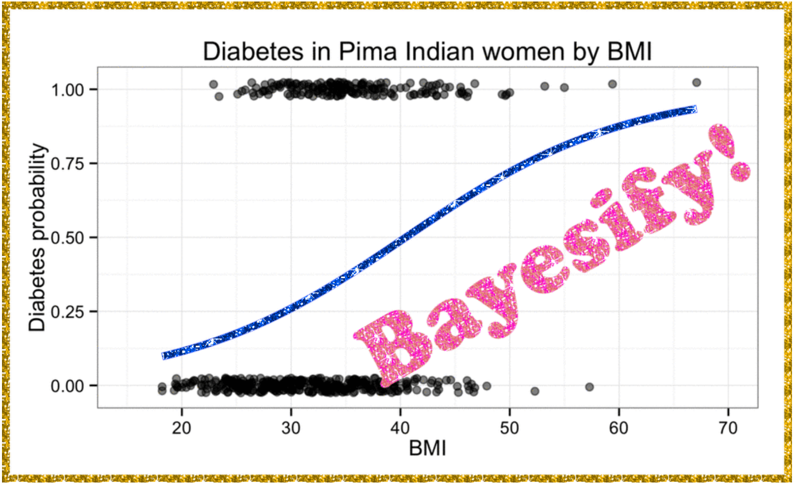
$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$$

Inferential goal: compute posterior $\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})$

Holmes & Held (H&H) set out to take regression models for categorical outcomes and ...



Holmes & Held (H&H) set out to take regression models for categorical outcomes and ...

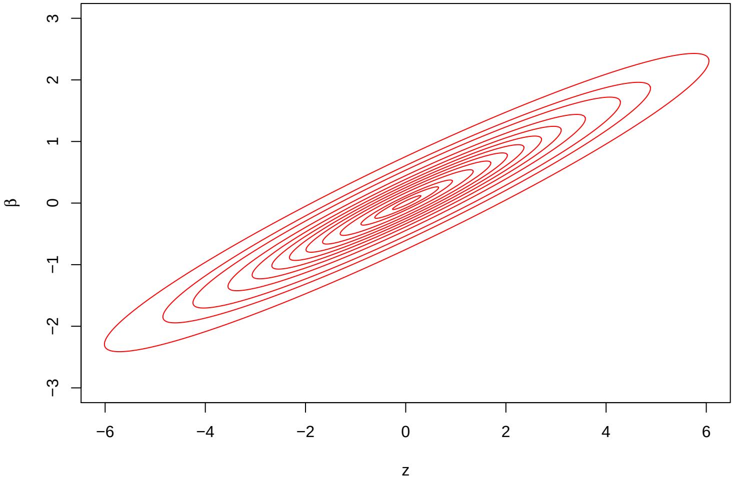


Why is logistic regression hard to Bayesify?

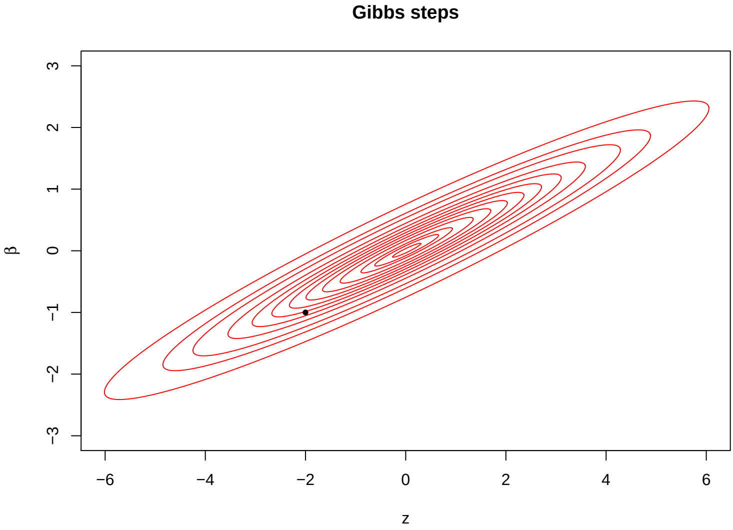
- Maximum likelihood not that easy either!
  - Fit using iterative methods
  - Asymptotics sidestep unknown finite sample distributions
- No conjugate priors ☹
- Most previous approaches involve Metropolis-Hastings and need tuning, or otherwise rely on accept-reject steps (e.g. Gamerman, 1997; Chen & Dey, 1998)
- Adaptive-rejection sampling (Dellaportas & Smith, 1993) only updates individual coefficients, resulting in poor mixing when coefficients are correlated

What we would like: **automatic and efficient Bayesian inference**

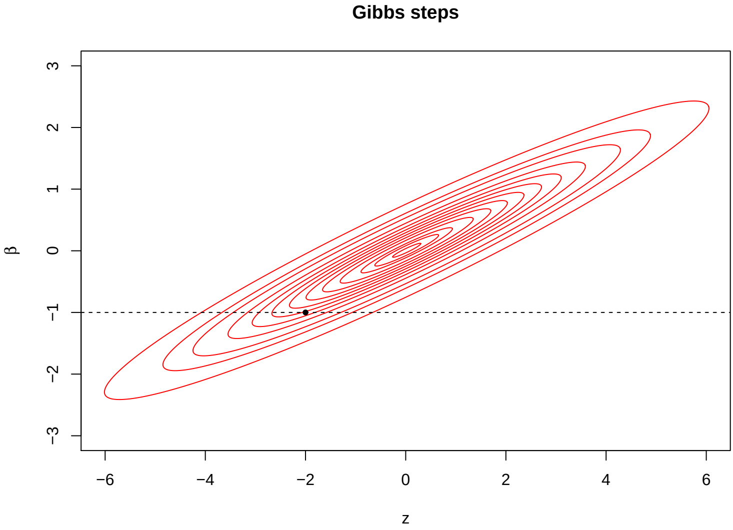
Mixing demonstration



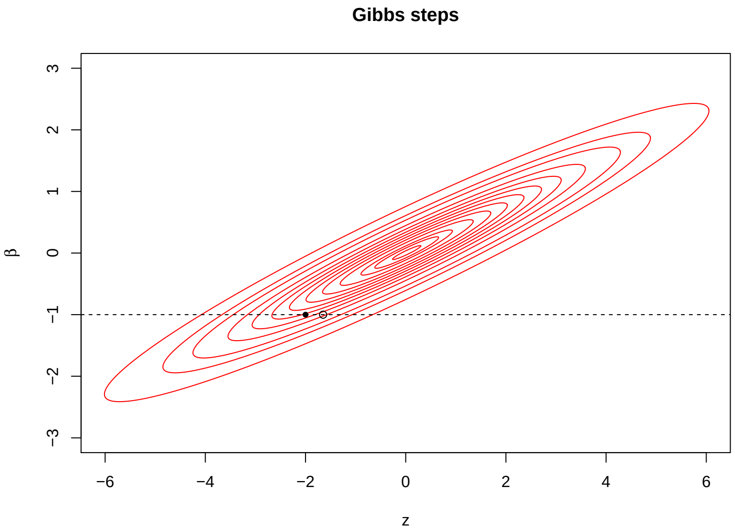
Mixing demonstration



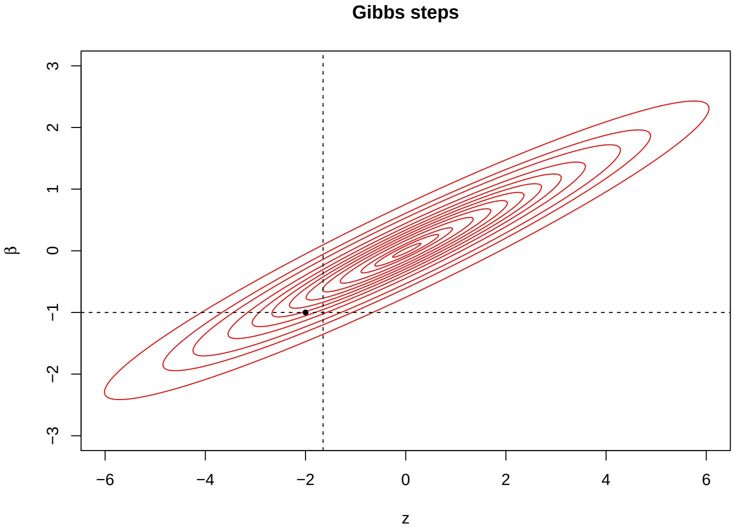
Mixing demonstration



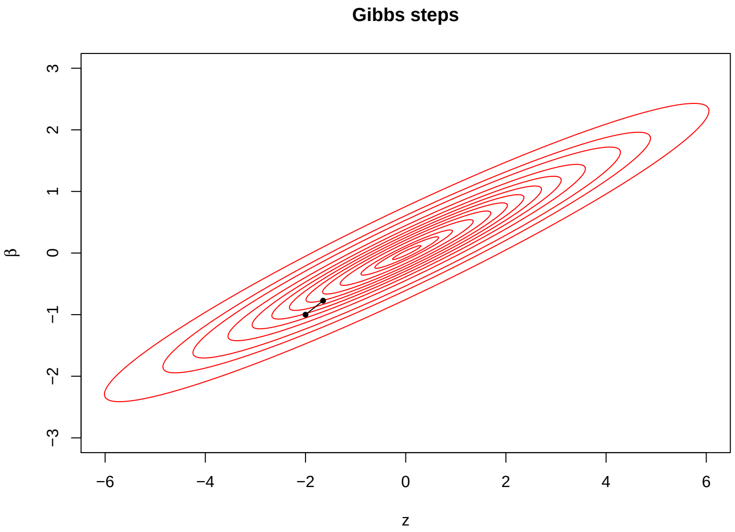
Mixing demonstration



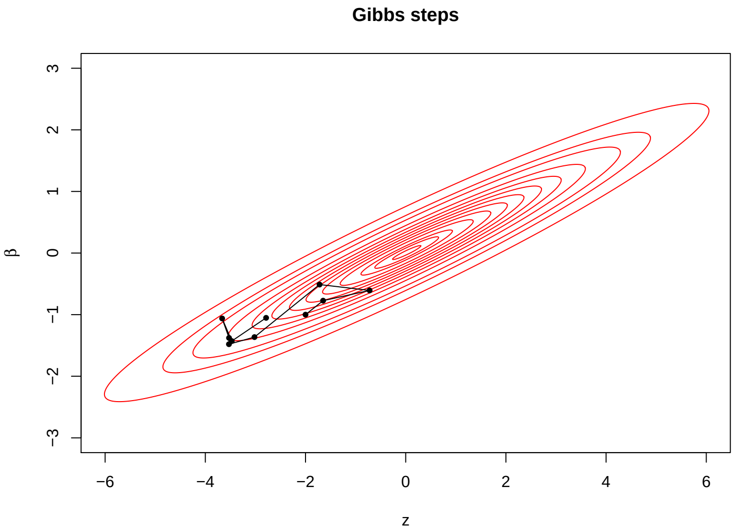
Mixing demonstration



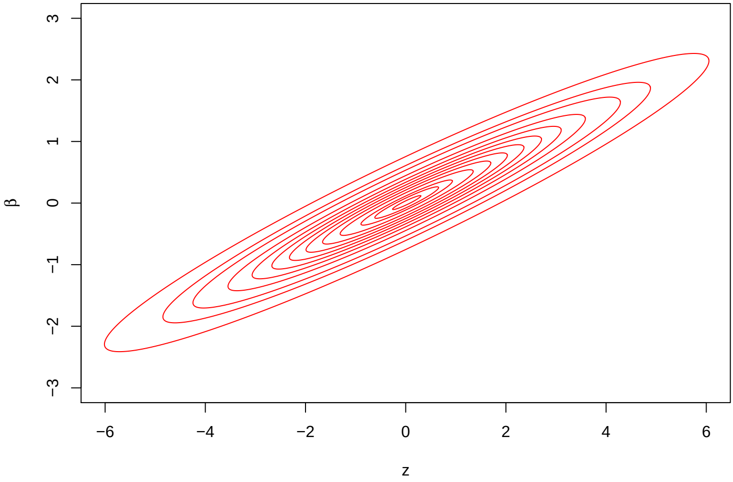
Mixing demonstration



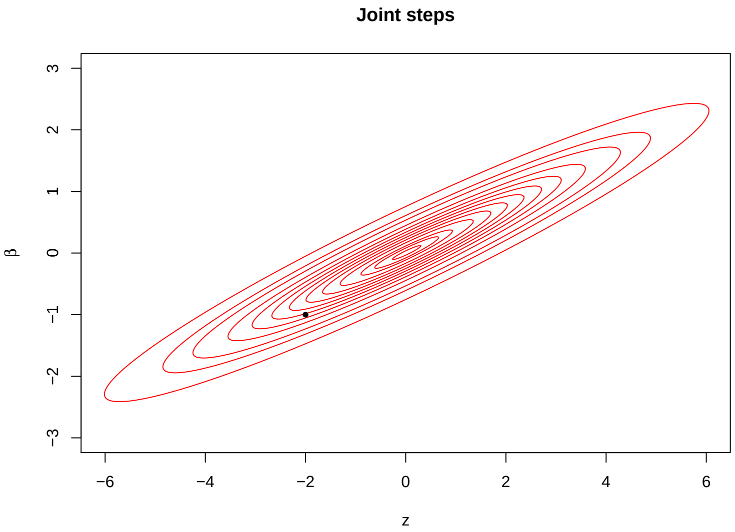
Mixing demonstration



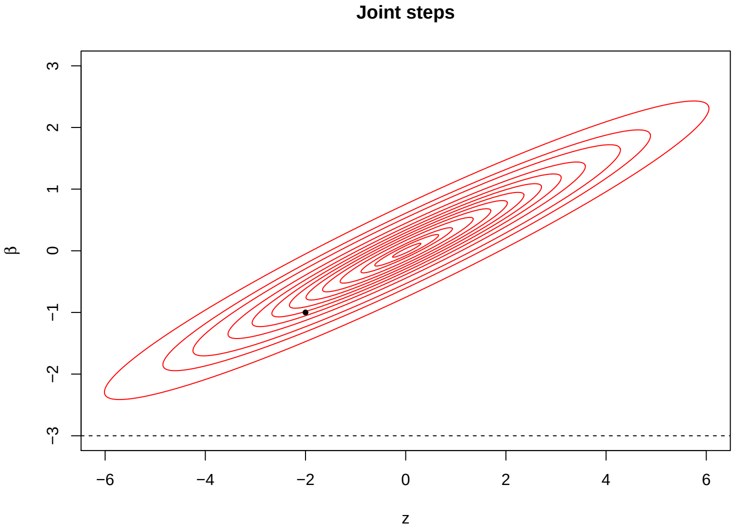
Mixing demonstration



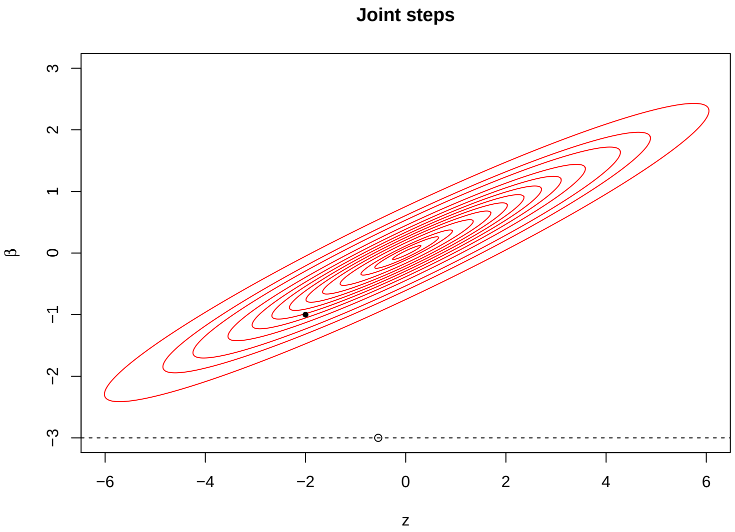
Mixing demonstration



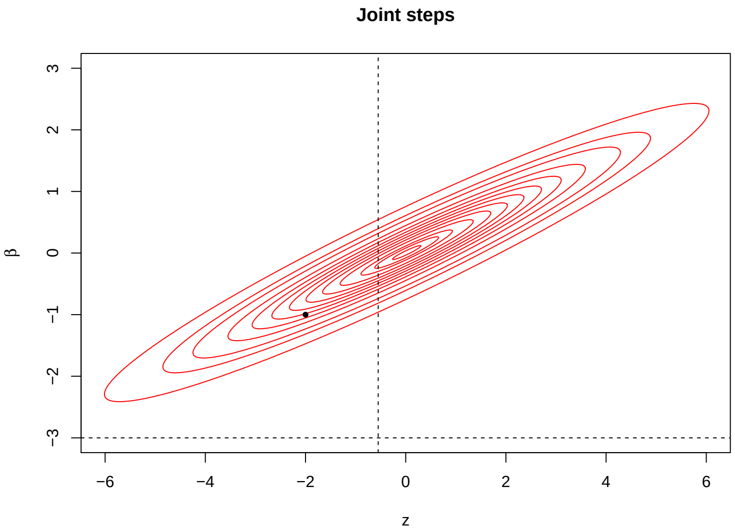
Mixing demonstration



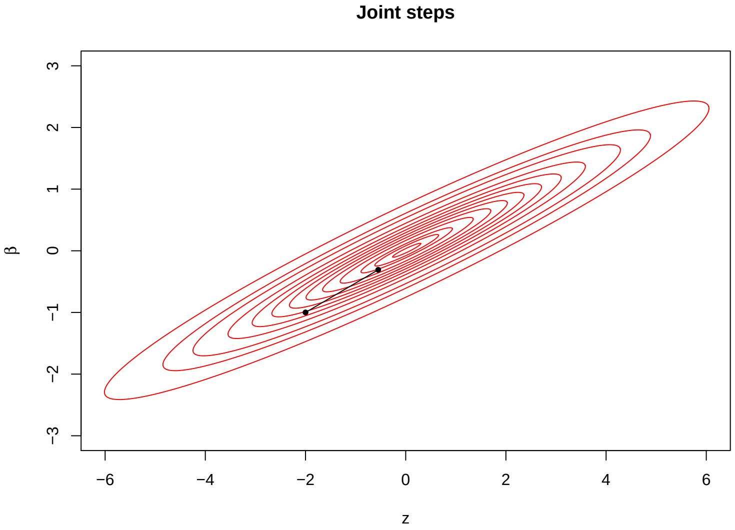
Mixing demonstration



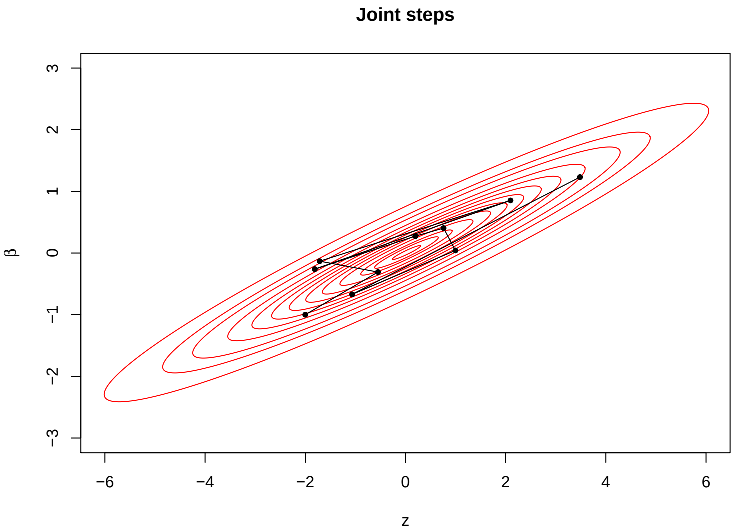
Mixing demonstration



Mixing demonstration



Mixing demonstration



H&H goals

H&H address four aspects of Bayesian inference for categorical data regression models:

(1) **Probit link**: use auxiliary variable method from Albert & Chib (A&C, 1993) to run MCMC automatically with Gibbs sampling, but with efficient joint updates

(2) **Logit link**: make auxiliary variable method and joint updating work with logistic regression

(3) **Model uncertainty**: extend methods to situations with uncertain covariate sets (e.g. Bayesian model averaging)

(4) **Polychotomous data**: extend methods to data with more than two outcomes

Probit regression

A&C auxiliary variable approach: introduce unobserved auxiliary variables z_i and re-write the probit model as

$$\begin{aligned} y_i &= 1_{[z_i > 0]} \\ z_i &= \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i \\ \epsilon_i &\sim N(0,1) \\ \boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}) \end{aligned}$$

Equivalent to probit model in standard framework:

$$\begin{aligned} p_i = P(z_i > 0 \mid \boldsymbol{\beta}) &= P(\mathbf{x}_i \boldsymbol{\beta} + \epsilon_i > 0 \mid \boldsymbol{\beta}) \\ = 1 - \Phi(-\mathbf{x}_i \boldsymbol{\beta}) &= \Phi(\mathbf{x}_i \boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i \boldsymbol{\beta}) \end{aligned}$$

Probit regression

From joint posterior, obtain nice conditional distributions of the parameters to simulate from in Gibbs steps:

$$\pi(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y}) \propto \underbrace{p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{z})}_{=p(\mathbf{y}|\mathbf{z})} p(\mathbf{z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}), \text{ so :}$$

- $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y}) \propto p(\mathbf{z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \underbrace{p(z_i \mid \boldsymbol{\beta})}_{N(\mathbf{x}_i\boldsymbol{\beta},1)}$

  If we use a normal prior for $\pi(\boldsymbol{\beta})$, then $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$ is also normal

- $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{z})p(\mathbf{z} \mid \boldsymbol{\beta})$

$$= \prod_{i=1}^{n} \underbrace{\left(1_{[z_i>0]}1_{[y_i=1]} + 1_{[z_i\leq 0]}1_{[y_i=0]}\right) \phi(z_i - \mathbf{x}_i\boldsymbol{\beta})}_{\pi(z_i|\boldsymbol{\beta},y_i) \cong \text{truncated normal}}$$

Smarter probit Gibbs

H&H improve mixing by updating $(\boldsymbol{\beta}, \mathbf{z})$ *jointly*: simulate from $\pi(\mathbf{z} \mid \mathbf{y})$, then from $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$. Assuming $\pi(\boldsymbol{\beta})$ normal:

$$\underbrace{\pi(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})}_{\text{(known form)}} = \underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})}_{\text{normal}} \pi(\mathbf{z} \mid \mathbf{y}) \text{ implies}$$

$$\pi(\mathbf{z} \mid \mathbf{y}) \sim \text{truncated multivariate normal}$$

Truncated multivariate normal hard to sample from, but univariate conditionals can be Gibbsed:

$$\pi(z_i \mid \mathbf{z}_{-i}, \mathbf{y}) \cong \begin{cases} N(m_i, v_i)\, 1_{[z_i > 0]} & \text{if } y_i = 1 \\ N(m_i, v_i)\, 1_{[z_i \leq 0]} & \text{if } y_i = 0 \end{cases}$$

where m_i and v_i are known (ugly) functions of $\mathbf{z}$, data, and prior

Logistic regression

So far: sampling the posterior for a Bayesian probit model can be done automatically and efficiently! ☺

Probit is a reasonable model for binary valued data, so why bother with a logit extension?

- Coefficients correspond to change in log odds
- Logit link has heavier tails than probit
- Probit link is not analytic and observations corresponding to extreme predicted probabilities can have numerical issues

From probit to logit

How to extend auxiliary variables to logistic regression?

$$
\begin{aligned}
y_i &= 1_{[z_i>0]} \\
z_i &= \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \\
\epsilon_i &\sim N(0, \lambda_i) \\
\lambda_i &= (2\psi_i)^2,\ \psi_i \sim KS \\
\boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta})
\end{aligned}
$$

Equivalent to logit model because ϵ_i has a logistic distribution (Andrews & Mallows, 1974) and CDF of logistic is expit function:

$$
\begin{aligned}
p_i = P(z_i > 0 \mid \boldsymbol{\beta}) &= P(\epsilon_i > -\mathbf{x}_i\boldsymbol{\beta} \mid \boldsymbol{\beta}) \\
&= 1 - \text{expit}(-\mathbf{x}_i\boldsymbol{\beta}) = \text{expit}(\mathbf{x}_i\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i\boldsymbol{\beta})
\end{aligned}
$$

Logistic Gibbs

In similar fashion to probit model, simulate from posterior conditionals:

$$\pi(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda} \mid \mathbf{y}) \propto \underbrace{p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda})}_{=p(\mathbf{y}|\mathbf{z})} p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \pi(\boldsymbol{\beta})$$

$$\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y}) \propto p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) \pi(\boldsymbol{\beta}) \cong \text{ normal if } \pi(\boldsymbol{\beta}) \text{ normal}$$

$$\pi(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) \cong \text{ indep. truncated normals}$$

$$\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{y}) \propto p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \cong \text{ indep. normal} \times \text{KS}^2$$

This last conditional distribution is non-standard, but easy to simulate from (no tuning needed)

Smarter logistic Gibbs

Joint updates for logistic to speed up mixing? A couple of possibilities:

(A) $\pi(\mathbf{z}, \boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})}_{\text{truncated logistic}} \underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z})}_{\text{rejection}}$ followed by $\underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda})}_{\text{normal}}$

(B) $\pi(\boldsymbol{\beta}, \mathbf{z} \mid \boldsymbol{\lambda}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\lambda}, \mathbf{y})}_{\text{truncated normal}} \underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda})}_{\text{normal}}$ followed by $\underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z})}_{\text{rejection}}$

Next time, aspirationally

- Performance of joint updating scheme for probit regression
- Performance of two joint updating schemes for logistic regression
- Auxiliary variable approaches under model uncertainty
- Auxiliary variable approaches with polychotomous outcomes
- What's happened since H&H 2006? (Go to the James Scott seminar next Thursday!)