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 #2

April 29, 2014

Categorical data setup

Classical framework with binary responses:

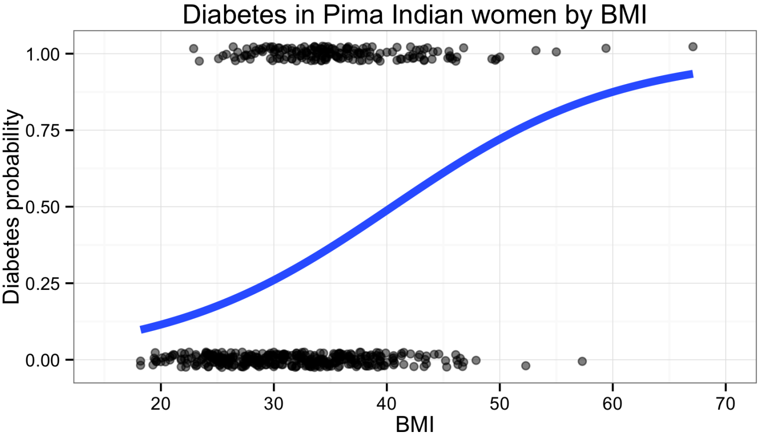
$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(p_i) \\
p_i &= g^{-1}(\eta_i),\ g^{-1} : \mathbb{R} \to (0,1) \\
\eta_i &= \mathbf{x}_i \boldsymbol{\beta},\ i = 1, \dots, n \\
\mathbf{x}_i &= \begin{pmatrix} x_{i1} & \dots & x_{ip} \end{pmatrix} \\
\boldsymbol{\beta} &= \begin{pmatrix} \beta_1 & \dots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

Put a prior on the unknown coefficients:

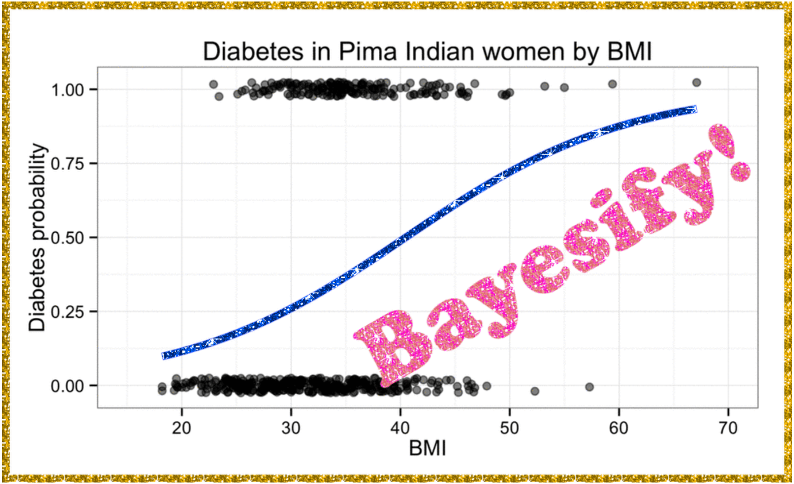
$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$$

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 ...

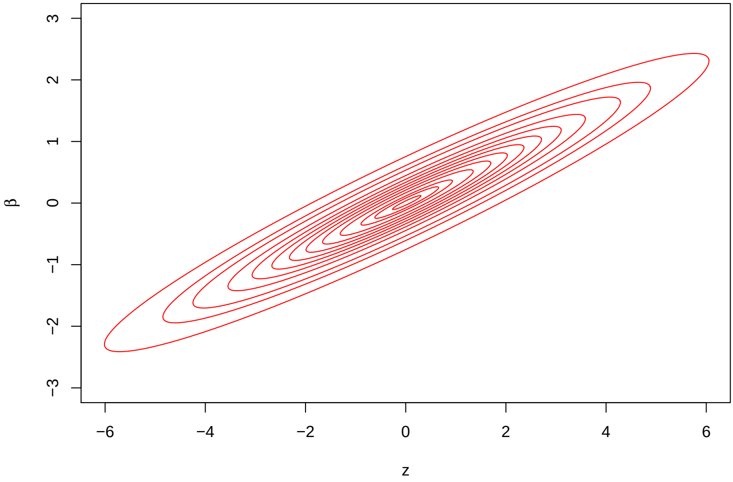


Why is logistic regression hard to Bayesify?

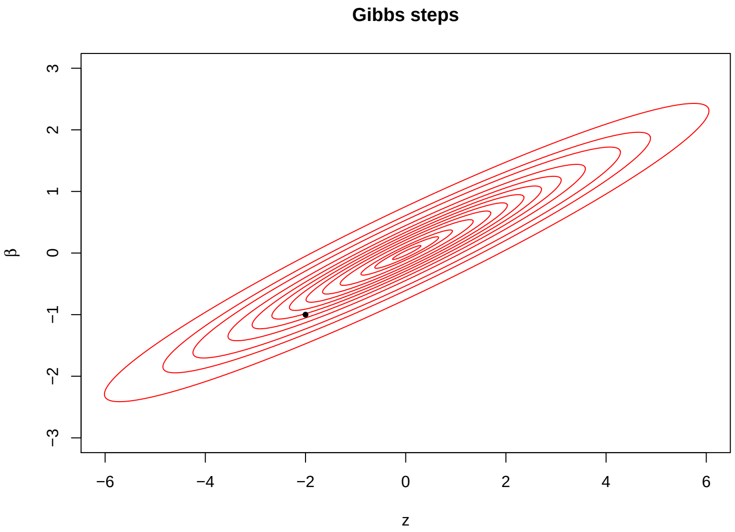
- Maximum likelihood not that easy either!
  - Fit using iterative methods
  - Asymptotics sidestep unknown finite sample distributions
- No conjugate priors ☹
- Most previous approaches involve Metropolis-Hastings and need tuning, or otherwise rely on accept-reject steps (e.g. Gamerman, 1997; Chen & 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**

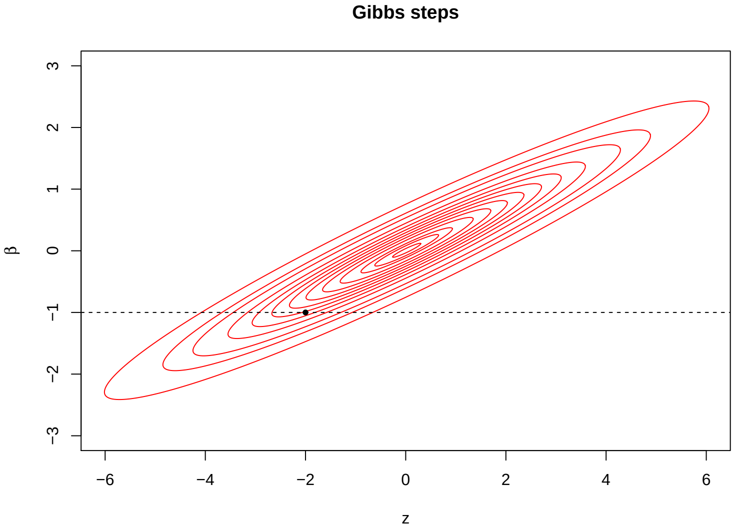
Mixing demonstration



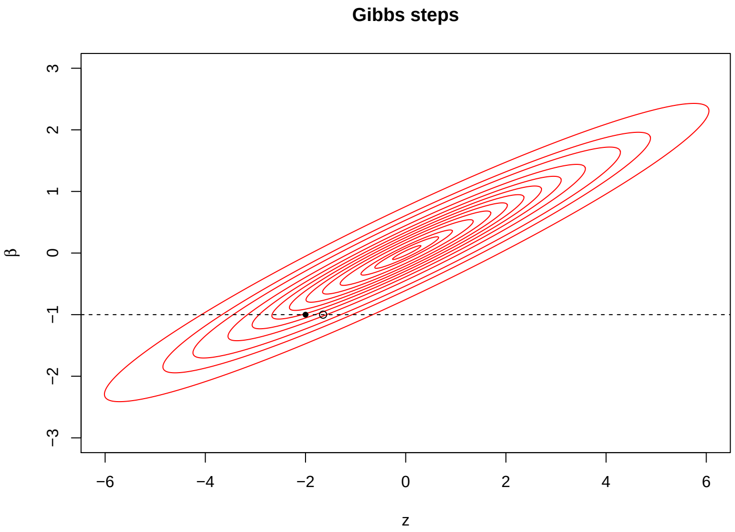
Mixing demonstration



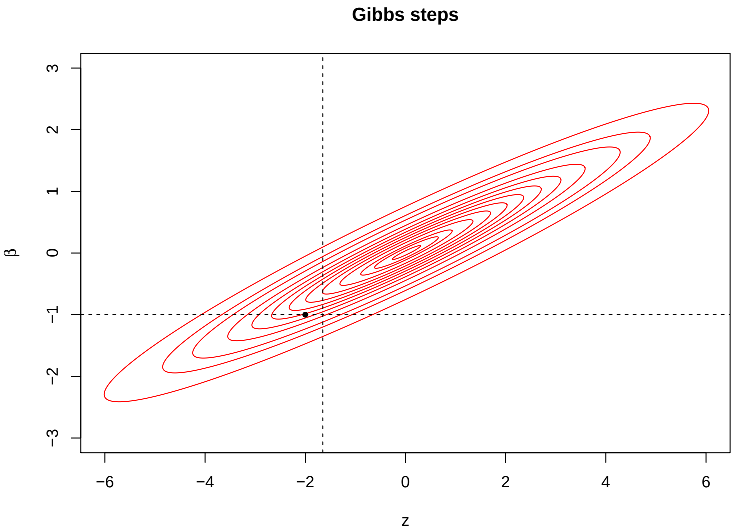
Mixing demonstration



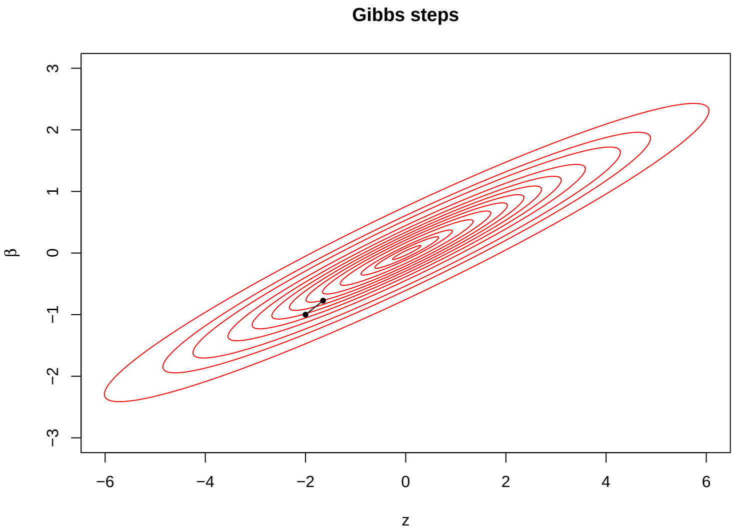
Mixing demonstration



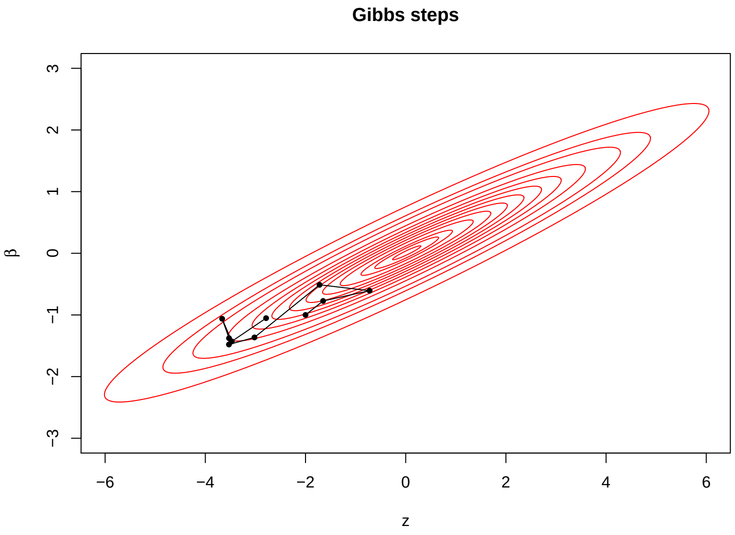
Mixing demonstration



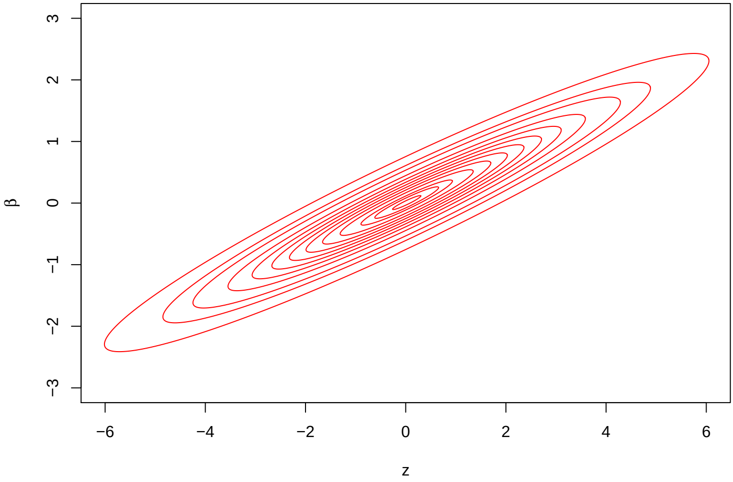
Mixing demonstration



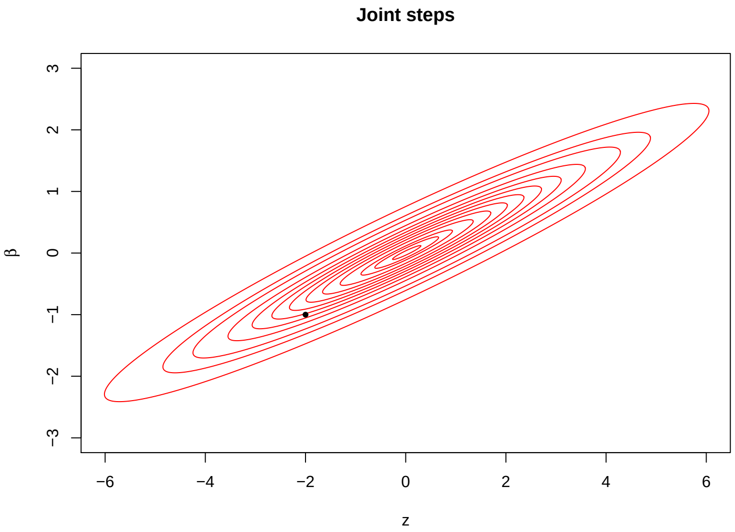
Mixing demonstration



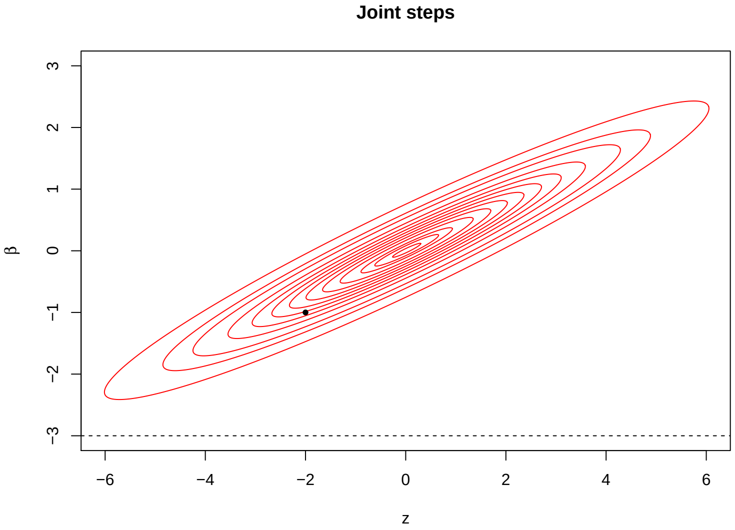
Mixing demonstration



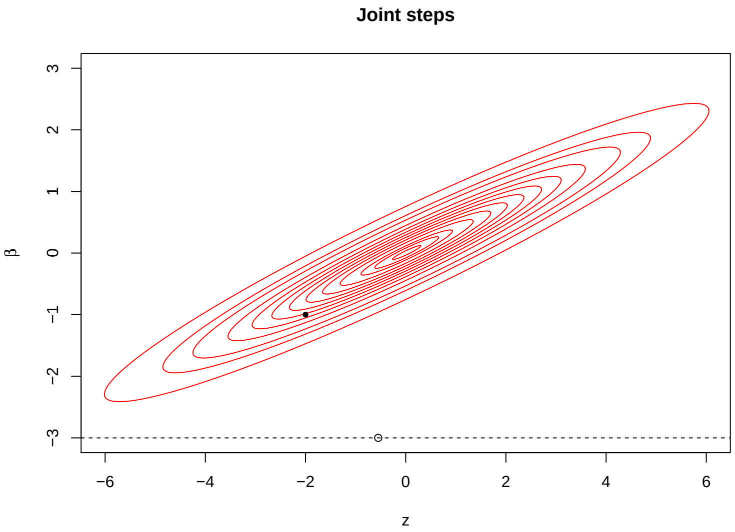
Mixing demonstration



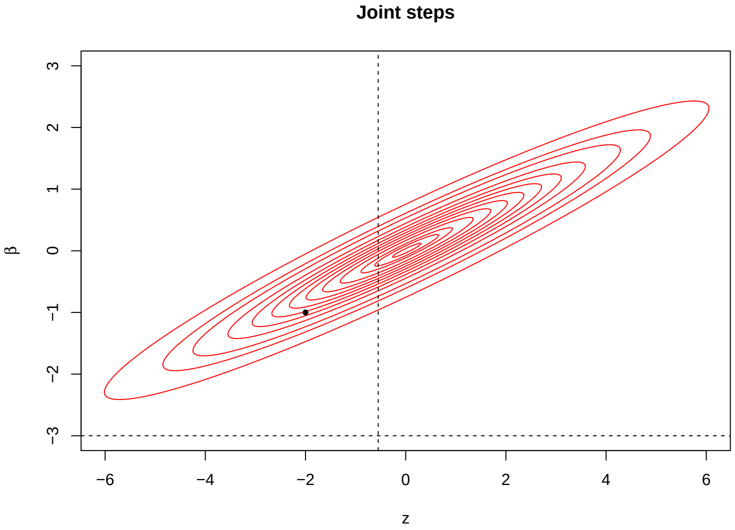
Mixing demonstration



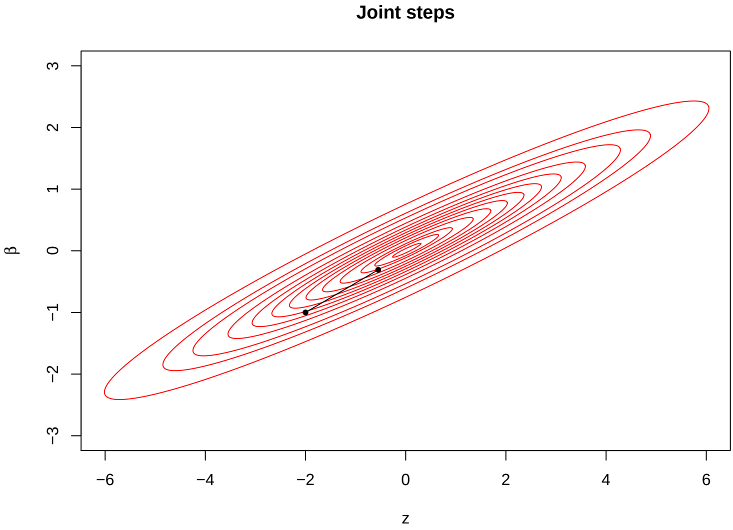
Mixing demonstration



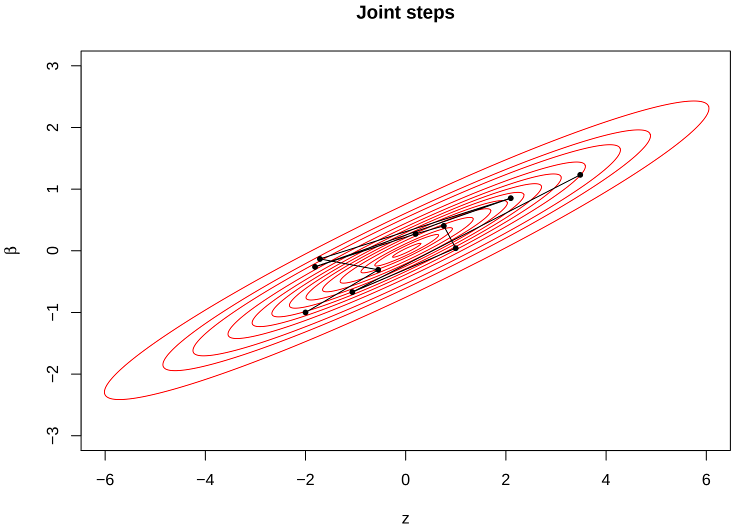
Mixing demonstration



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) **Polytomous 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

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

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 Gibbs steps (A&C)

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 sampling?

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 directly, but univariate conditionals can be Gibbsed:

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

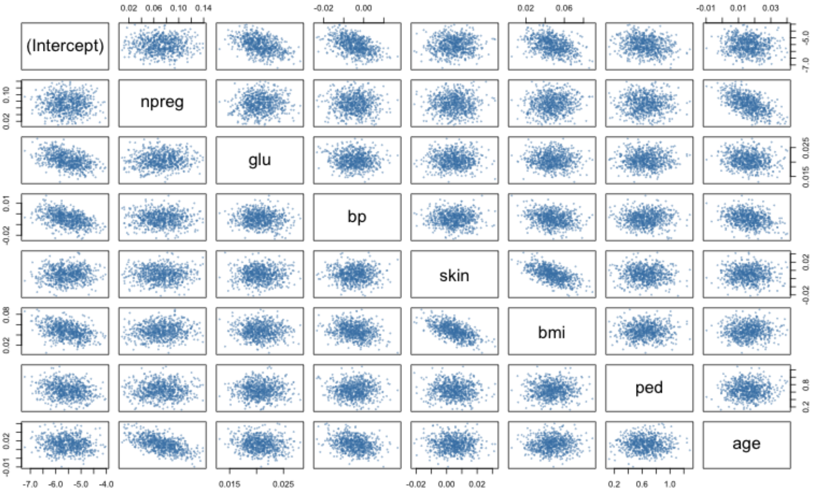
where m_i and v_i are known (ugly) functions of $\mathbf{z}$, data, and prior

Test data

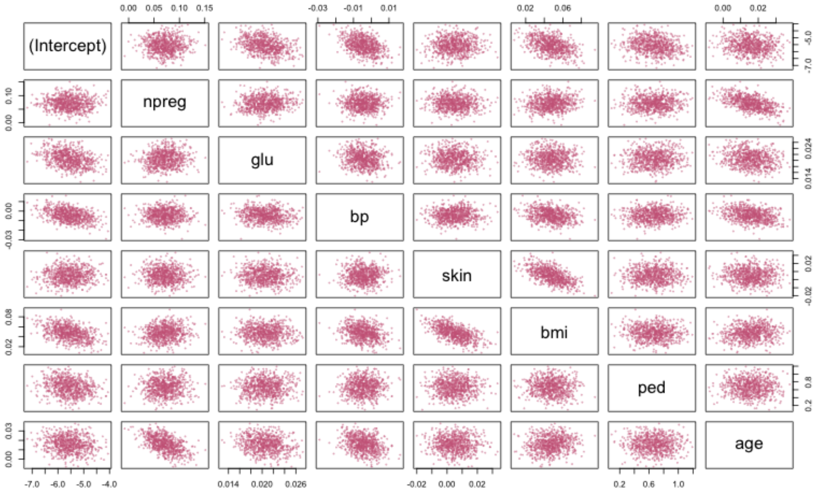
H&H analyze several stock datasets with binary outcomes:

- Pima Indian data ($n = 532, p = 8$): outcome is diabetes; covariates include BMI, age, number of pregnancies
- Australian credit data ($n = 690, p = 14$): outcome is credit approval; 14 generic covariates
- Heart disease data ($n = 270, p = 13$): outcome is heart disease; covariates include age, sex, blood pressure, chest pain type
- German credit data ($n = 1000, p = 24$): outcome is good vs. bad credit risk; covariates include checking account status, purpose of loan, gender and marital status

Example probit posterior: iterative sampling



Example probit posterior: joint sampling



Efficient Bayesian inference

How might we see if a MCMC sampling algorithm is efficient?

- Time elapsed to run M iterations
- Average update distance: measure mixing with

$$\frac{1}{M-1}\sum_{i=1}^{M-1}\|\boldsymbol{\beta}^{(i+1)}-\boldsymbol{\beta}^{(i)}\|$$

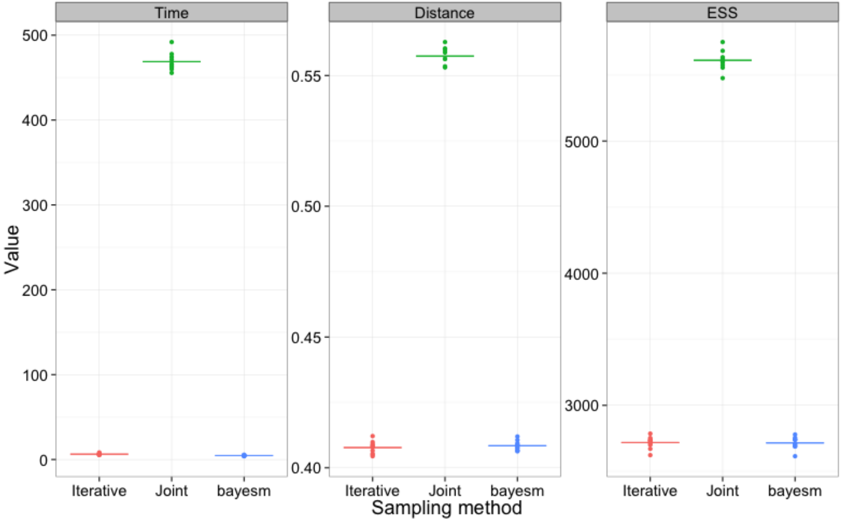
- Effective sample size (ESS) for a single parameter:

$$\text{ESS}=\frac{M}{1+2\sum_{k=1}^{\infty}\rho(k)}$$

  where $\rho(k) =$ monotone sample autocorrelation at lag k (Kass et al, 1998)

Testing procedure: compute these metrics on each of 10 runs of $M =$10,000 iterations per run (discard 1,000 burn-in)

Probit performance: absolute



From probit to logit

Extend auxiliary variables to logistic regression with another level for variance of the error terms:

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

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:

$$\pi(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda} \mid \mathbf{y}) \propto \underbrace{p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda})}_{=p(\mathbf{y}|\mathbf{z}) \text{ truncators}} \underbrace{p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda})}_{\text{indep. normal}} \underbrace{p(\boldsymbol{\lambda})}_{\text{KS}^2} \underbrace{\pi(\boldsymbol{\beta})}_{\text{normal}}$$

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

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

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

Last conditional distribution is non-standard, but can be simulated using rejection sampling with Generalized Inverse Gaussian proposals and alternating series representation ("squeezing")

Joint updates for mixing

H&H propose using factorizations of the joint posterior for updates.

- Probit: simulate from $\pi(\mathbf{z} \mid \mathbf{y})$, then from $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$

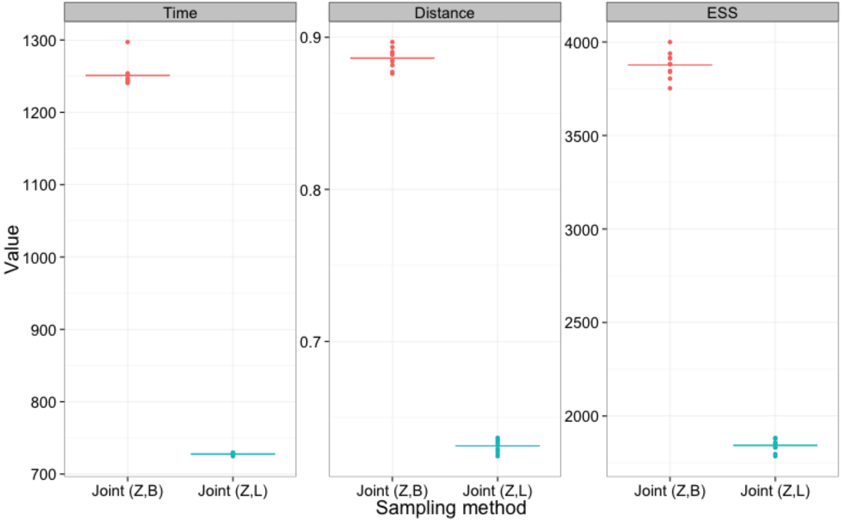
$$\pi(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \mathbf{y})}_{\text{truncated multivariate normal}} \quad \underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})}_{\text{normal}}$$

- Logistic: a couple of possibilities

(A) $\pi(\mathbf{z}, \boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})}_{\text{truncated ind logistic}} \quad \underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z})}_{\text{normal} \times \text{KS}^2}$, then $\underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda})}_{\text{normal}}$

(B) $\pi(\boldsymbol{\beta}, \mathbf{z} \mid \boldsymbol{\lambda}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\lambda}, \mathbf{y})}_{\text{truncated mv normal}} \quad \underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda})}_{\text{normal}}$, then $\underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z})}_{\text{normal} \times \text{KS}^2}$

Logistic performance: absolute



Model uncertainty

Suppose we have our set of p covariates but don't know which to include in our logistic regression model.

An approach: yet **more latent variables**

$$\gamma_j = \begin{cases} 1 & \text{if } \beta_j \text{ in model} \\ 0 & \text{if } \beta_j \text{ not in model} \end{cases}, \; j = 1, \ldots, p$$

Now, we condition $\boldsymbol{\beta}$ on $\boldsymbol{\gamma}$ so $z_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$ becomes

$$z_i = \mathbf{x}_{i\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} + \epsilon_i = \sum_{j=1}^{p} x_{ij\boldsymbol{\gamma}} \beta_j \gamma_j + \epsilon_i$$

Then: estimate $\pi(\gamma_j = 1 \mid \mathbf{y})$ (among other interesting quantities)

Updating scheme

Posterior:

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

Update sets of coefficients with blocked Gibbs iterations:

(1) $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y}) \propto \underbrace{p(\mathbf{z} \mid \boldsymbol{\beta}, \gamma, \boldsymbol{\lambda})}_{N(\mathbf{x}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\Lambda}_{\boldsymbol{\gamma}})} \underbrace{\pi(\boldsymbol{\beta} \mid \gamma)}_{N(\mathbf{b}_{\boldsymbol{\gamma}}, \mathbf{v}_{\boldsymbol{\gamma}})} \pi(\boldsymbol{\gamma})$ using M-H

(2) $\pi(\mathbf{z}, \boldsymbol{\lambda} \mid \gamma, \boldsymbol{\beta}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\beta}, \gamma, \mathbf{y})}_{\text{truncated logistic}} \underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \gamma, \mathbf{z})}_{\text{normal} \times \text{KS}^2}$

Note that we update $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ simultaneously and jump dimensions – much harder to do with iterative sampling

Metropolis-Hastings step

Target density:

$$\pi(\gamma, \boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y}) \propto \underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \gamma, \boldsymbol{\lambda}, \mathbf{y})}_{N(\mathbf{B}_\gamma, \mathbf{V}_\gamma)} \pi(\gamma)$$

with $\mathbf{B}_\gamma, \mathbf{V}_\gamma$ determined by $\gamma, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{b}, \mathbf{v}, \mathbf{x}$

(i) Given current $(\gamma, \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda})$, propose from

$$Q(\gamma^*, \boldsymbol{\beta}^* \mid \gamma, \boldsymbol{\beta}) = \underbrace{q(\gamma^* \mid \gamma)}_{\text{proposal density}} \underbrace{\pi(\boldsymbol{\beta}^* \mid \mathbf{z}, \gamma^*, \boldsymbol{\lambda}, \mathbf{y})}_{N(\mathbf{B}_{\gamma^*}, \mathbf{V}_{\gamma^*})}$$

(ii) Accept $(\gamma^*, \boldsymbol{\beta}^*)$ as update with probability

$$\alpha = \min\left\{1, \frac{|\mathbf{V}_{\gamma^*}|^{1/2}|\mathbf{v}_\gamma|^{1/2}\exp(0.5\mathbf{B}_{\gamma^*}^T\mathbf{V}_{\gamma^*}^{-1}\mathbf{B}_{\gamma^*})\pi(\gamma^*)q(\gamma \mid \gamma^*)}{|\mathbf{V}_\gamma|^{1/2}|\mathbf{v}_{\gamma^*}|^{1/2}\exp(0.5\mathbf{B}_\gamma^T\mathbf{V}_\gamma^{-1}\mathbf{B}_\gamma)\pi(\gamma)q(\gamma^* \mid \gamma)}\right\}$$

(iii) Otherwise stay in current state of $(\gamma, \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\lambda})$

From dichotomous to polytomous

Generalize the logistic regression model for classification problems by allowing unordered outcomes $\{1, 2, \ldots, Q\}$ instead of $\{0, 1\}$:

$$
\begin{aligned}
y_i &\sim \text{Multinomial}(\theta_{i1}, \ldots, \theta_{iQ}) \\
\theta_{ij} &= \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{\sum_{k=1}^{Q} \exp(\mathbf{x}_i \boldsymbol{\beta}_k)} \\
\boldsymbol{\beta}_Q &= \mathbf{0} \text{ for identifiability}
\end{aligned}
$$

Polytomous sampling

Conditional likelihood has form of binary logistic regression:

$$L(\boldsymbol{\beta}_j \mid \mathbf{y}, \boldsymbol{\beta}_{-j}) \propto \prod_{i=1}^{n} \left(\underbrace{\frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j - C_{ij})}{1 + \exp(\mathbf{x}_i \boldsymbol{\beta}_j - C_{ij})}}_{\eta_{ij}} \right)^{[y_i = j]} \cdot (1 - \eta_{ij})^{[y_i \neq j]}$$

$$C_{ij} = \sum_{k \neq j} \log \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$$

so in Bayesian framework bringing in priors and auxiliary variables, we can Gibbs over each of the $Q-1$ classes and treat each using either of the logistic regression sampling schemes

To do/lingering concerns

- More simulations: additional datasets, iterative updates for logistic, model uncertainty, polytomous regression
- Numerical and speed issues with rejection sampler for conditional distribution of λ in logistic models
- Deriving acceptance rate for M-H steps under model uncertainty
- Scope issue: how much time/effort to devote to discussing later work? (e.g. Pólya-Gamma model by Polson, Scott, and Windle, or refinements by Frühwirth-Schnatter, Frühwirth, Rue)