Bayesian auxiliary variable models for binary and multinomial regression

(*Bayesian Analysis*, 2006)

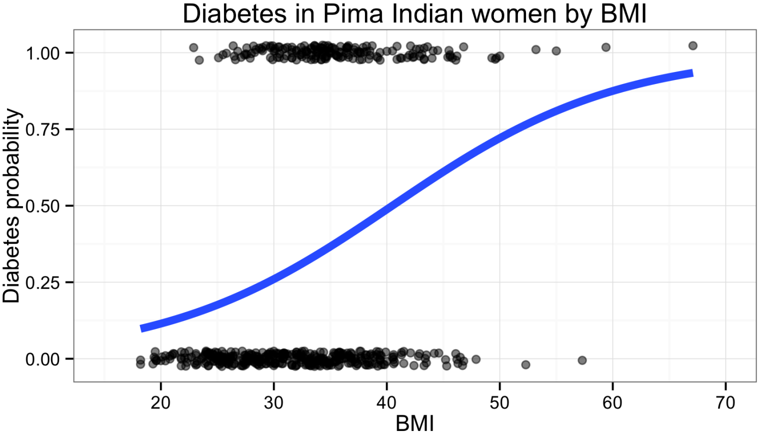
Authors: Chris Holmes
Leonhard Held

As interpreted by: Rebecca Ferrell

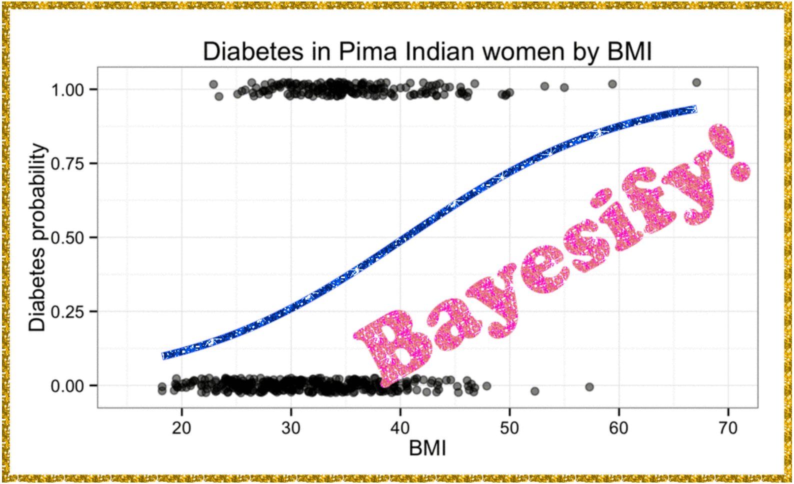
UW Statistics 572, Final Talk

May 27, 2014

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



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



Outline

- Introduction
  - Intro to probit and logistic regression in Bayesian context
  - Quick overview of the Gibbs sampler
- Probit regression
  - Review popular way of doing Bayesian probit regression from 1993 by Albert & Chib (A&C)
  - Compare Holmes & Held (H&H) probit approach with A&C
- Logistic regression
  - H&H's modifications to make ideas work for logistic regression
  - Empirical performance of sampling strategies
- Discussion
  - ~~Extension to model uncertainty~~ (no time!)
  - ~~Extension to multiple outcomes~~ (no time!)
  - Concluding thoughts

Binary data setup

Classical framework with n binary responses y_i and covariates $\mathbf{x}_i$:

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

Put a prior on the unknown coefficients:

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

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

Why are binary regression models hard to Bayesify?

- No conjugate priors – will need to use MCMC sampling
  - (Max. likelihood needs iterative methods, asymptotics)
- Previous approaches involve sampling from an approximation to the posterior, need tuning, or otherwise rely on data-dependent 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

Wishlist for **automatic and efficient Bayesian inference**:

- MCMC samples from exact posterior distribution
- No tuning of proposal distributions or low accept-reject rates
- Reasonable mixing even with correlated parameters

Intro to Gibbs sampling

Setup: we don't know posterior distribution $\pi(\boldsymbol{\beta} \mid \mathbf{y})$, but do know each conditional posterior $\pi(\beta_i \mid \boldsymbol{\beta}_{-i}, \mathbf{y})$.

Gibbs sampling **iterates over conditional posteriors** to produce a sample from a Markov chain with stationary distribution $\pi(\boldsymbol{\beta} \mid \mathbf{y})$:

(1) Initialize $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \dots, \beta_p^{(0)})$

(2) Draw $\beta_1^{(1)} \sim \pi(\beta_1 \mid \boldsymbol{\beta}_{-1}^{(0)}, \mathbf{y})$

(3) Draw $\beta_2^{(1)} \sim \pi(\beta_2 \mid \beta_1^{(1)}, \boldsymbol{\beta}_{-\{1,2\}}^{(0)}, \mathbf{y}) \dots$

(4) ... Draw $\beta_p^{(1)} \sim \pi(\beta_p \mid \boldsymbol{\beta}_{-p}^{(1)}, \mathbf{y})$

(5) Done with sample observation $\boldsymbol{\beta}^{(1)}$, now repeat (2) - (4)

Combine Gibbs steps into blocks: e.g. if distribution of $\pi(\beta_1, \beta_2 \mid \boldsymbol{\beta}_{-\{1,2\}}, \mathbf{y})$ is available, can use in place of the individual conditionals in (2) and (3).

Probit regression

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

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

Equivalent to probit model with $y_i \sim \text{Bernoulli}\,(p_i = \Phi(\mathbf{x}_i \boldsymbol{\beta}))$:

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

Probit the A&C way: iterative Gibbs steps

From joint posterior, obtain nice block conditional distributions of the parameters to iterate through in Gibbs steps:

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

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

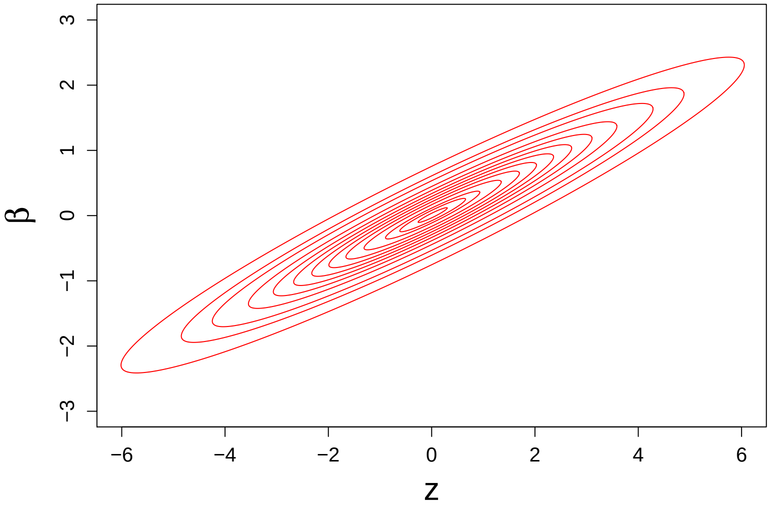
$=$ multivariate normal

(2) $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i \mid z_i) p(z_i \mid \boldsymbol{\beta})$

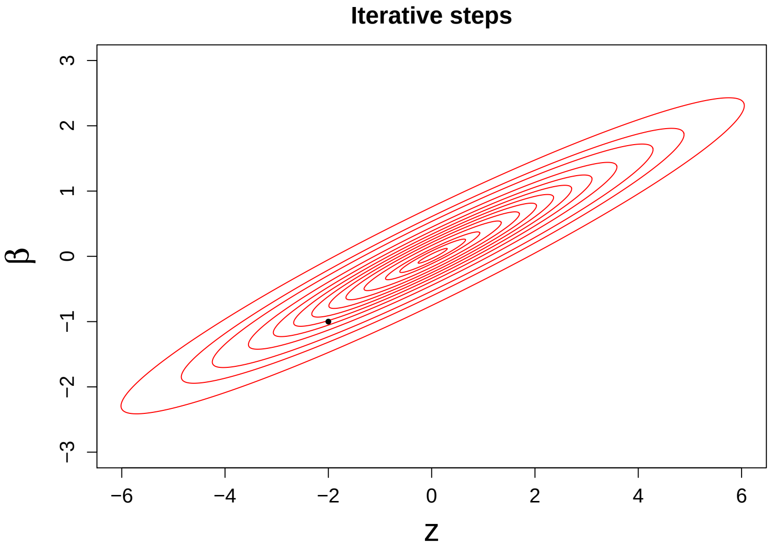
$$= \prod_{i=1}^{n} \underbrace{\left(1_{[z_i>0]} 1_{[y_i=1]} + 1_{[z_i \leq 0]} 1_{[y_i=0]}\right) \phi(z_i - \mathbf{x}_i \boldsymbol{\beta})}_{\pi(z_i|\boldsymbol{\beta},y_i) \cong \text{truncated normal}}$$

$=$ product of truncated univariate normals

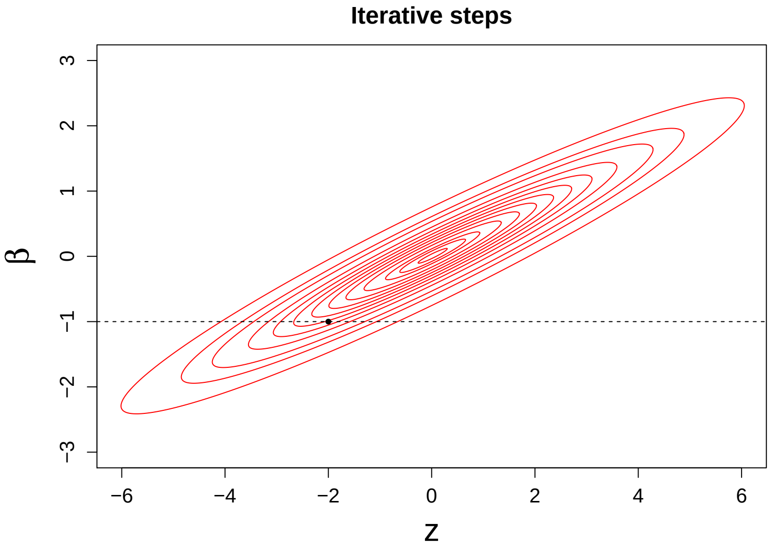
Mixing demonstration



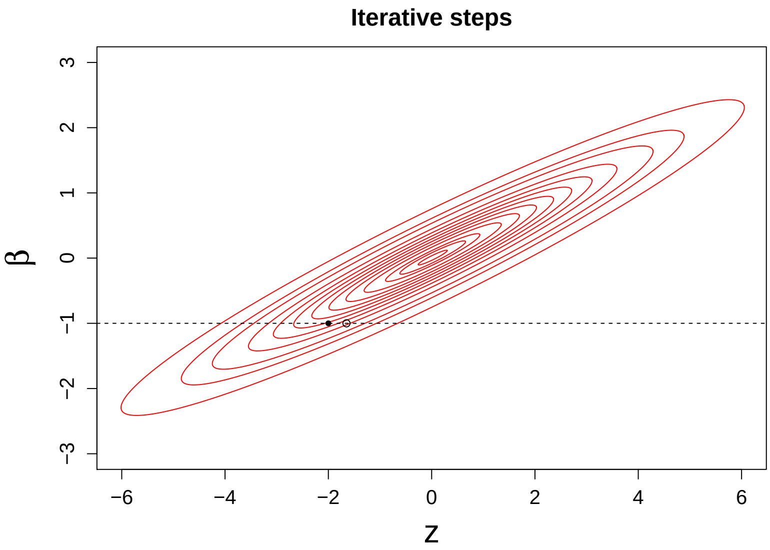
Mixing demonstration



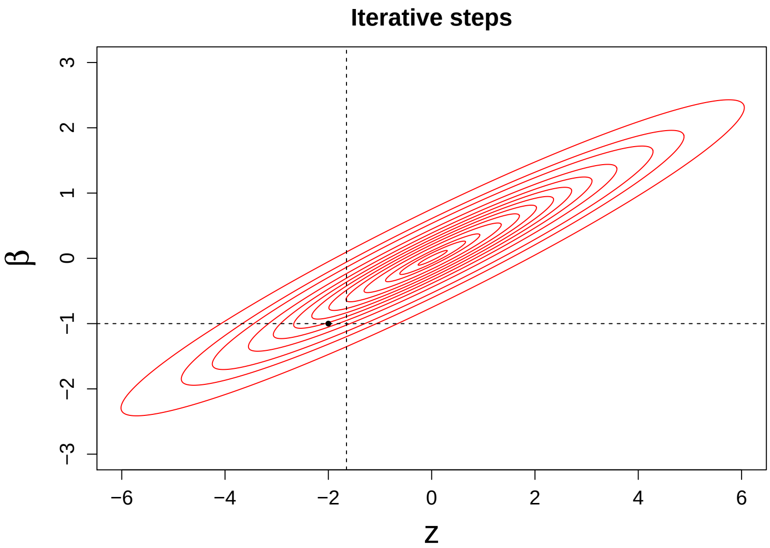
Mixing demonstration



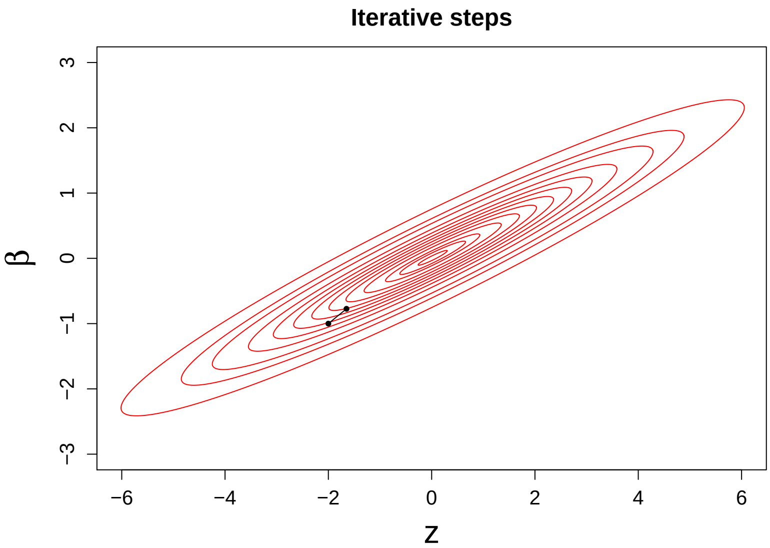
Mixing demonstration



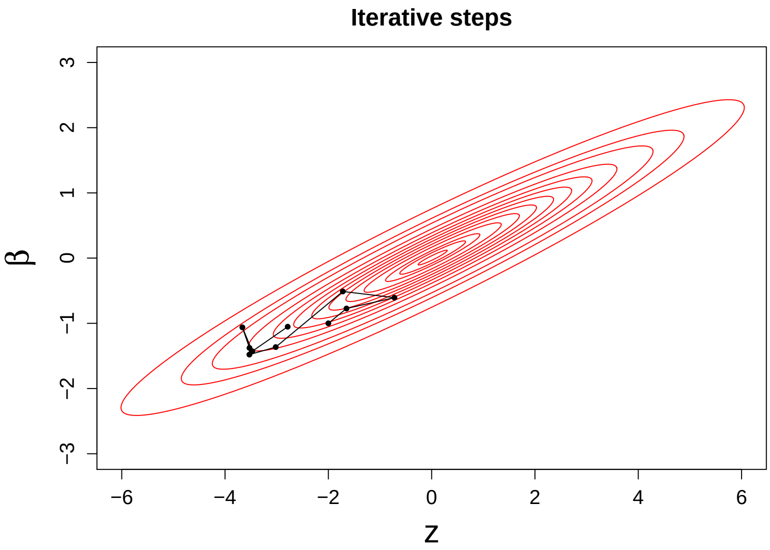
Mixing demonstration



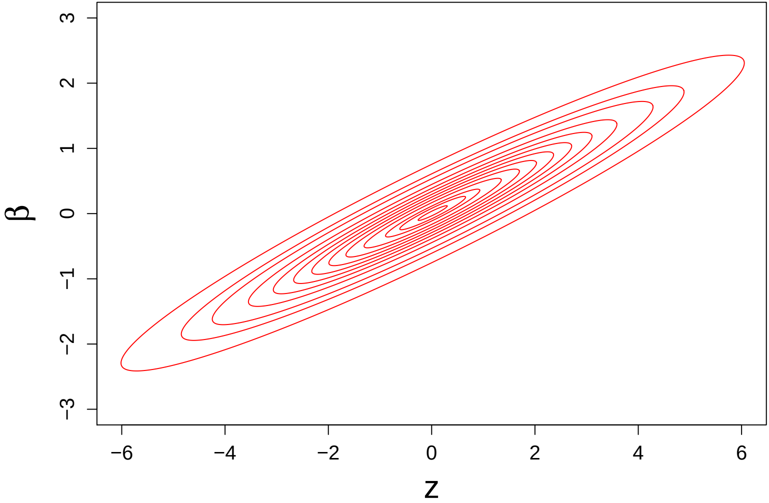
Mixing demonstration



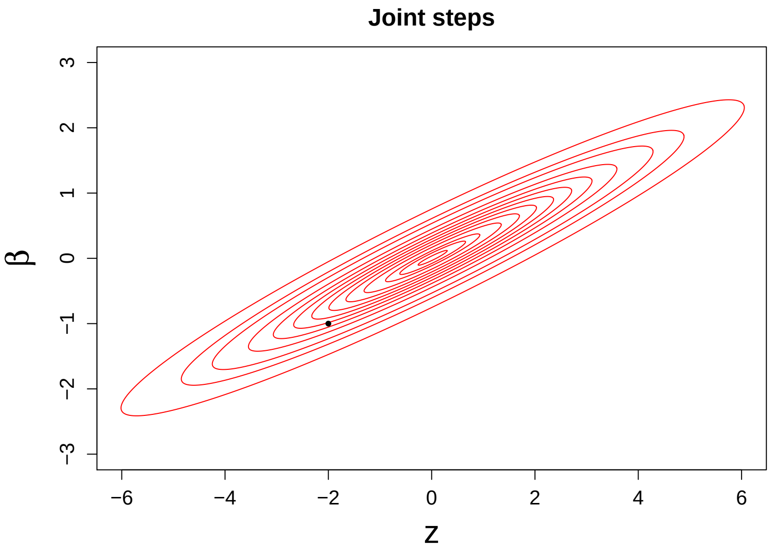
Mixing demonstration



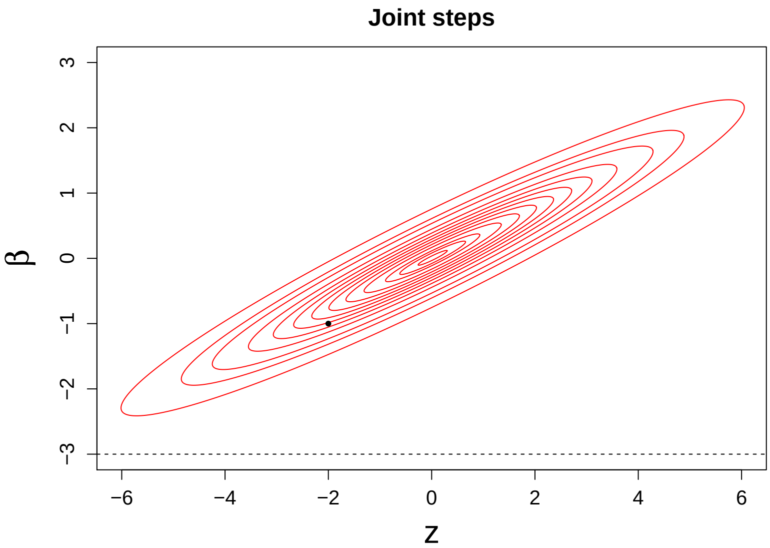
Mixing demonstration



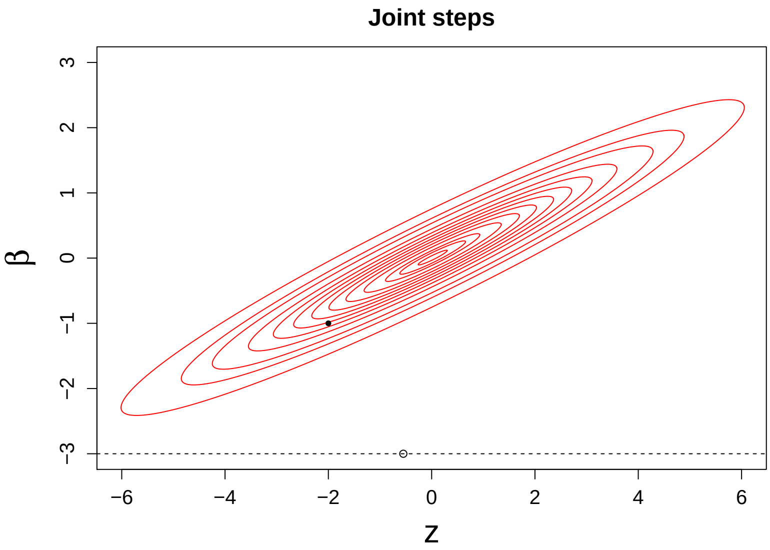
Mixing demonstration



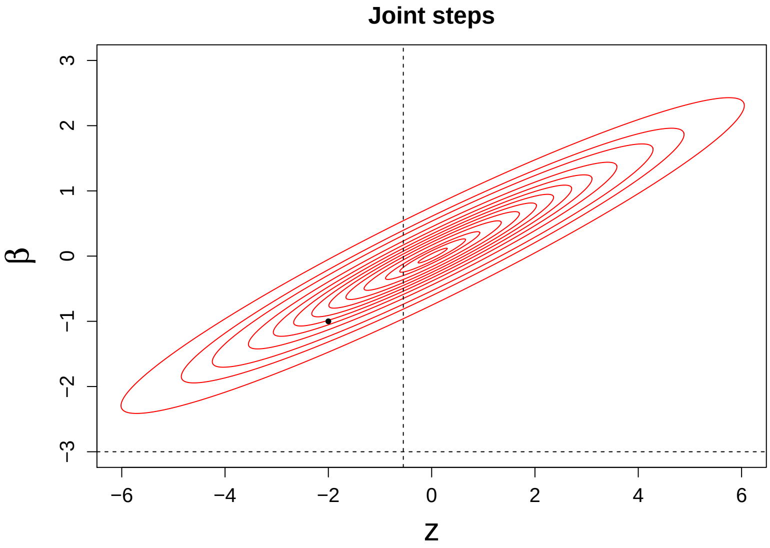
Mixing demonstration



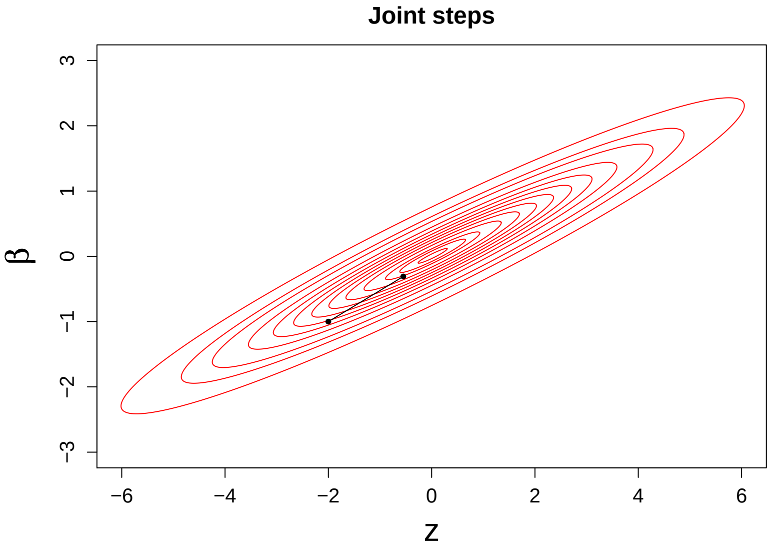
Mixing demonstration



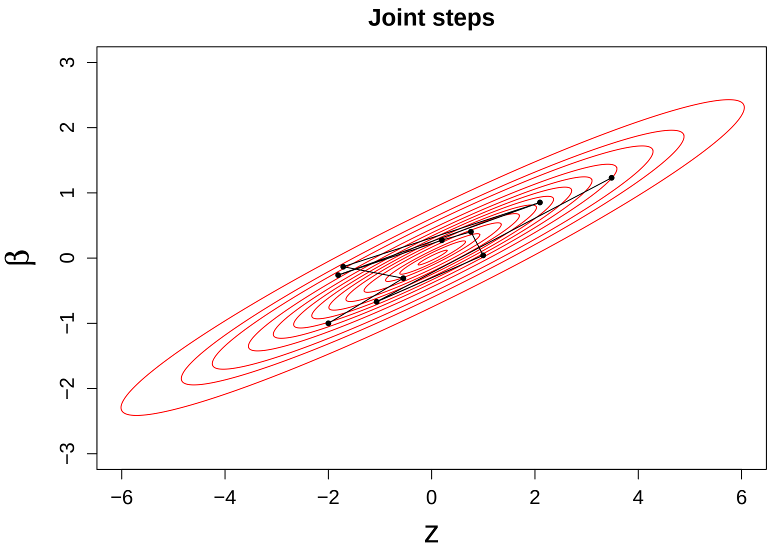
Mixing demonstration



Mixing demonstration



Mixing demonstration



Smarter Gibbs sampling for probit?

H&H improve mixing by updating $(\boldsymbol{\beta}, \mathbf{z})$ *jointly*: simulate from $\pi(\mathbf{z} \mid \mathbf{y})$, then from $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$. With $\pi(\boldsymbol{\beta})$ 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 very hard to sample from directly, 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 leave-one-out functions of $\mathbf{z}_{-i}$, data, and prior

Probit sampler comparison

Iterative updates from A&C:

- Iterate between block Gibbs updates $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})$, $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$
- $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y}) \sim n$ independent truncated normals with variance 1
- Blocking, independence need just two matrix updates per cycle, should run quickly — *implementation in H&H paper appears not to have exploited this for* $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})$

Joint updates from H&H:

- Iterate through n univariate Gibbs updates $\pi(z_i \mid \mathbf{z}_{-i}, \mathbf{y})$, then one block Gibbs update $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$
- $\pi(z_i \mid \mathbf{z}_{-i}, \mathbf{y}) \sim$ truncated normal with variance $v_i > 1$
- Can't do the z_i's all at once, need $n + 1$ matrix calculations per cycle — but maybe bigger variance can offset slowness through better mixing?

Efficient Bayesian inference

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

(1) **Time elapsed** to run M iterations

(2) **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)

(3) **Average update distance**: measure mixing with

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

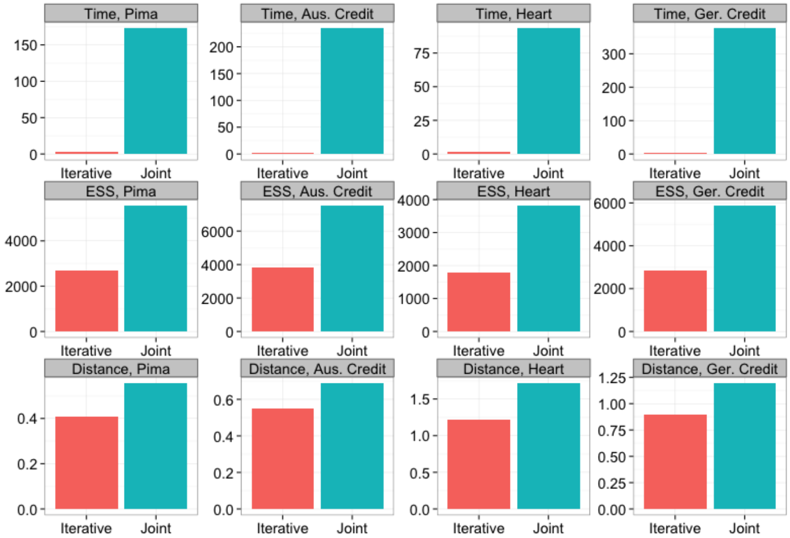
Testing procedure: compute these metrics on each of 10 runs of $M =$10,000 iterations per run (discard 1,000 burn-in)

Test data

H&H analyze several stock datasets with binary outcomes:

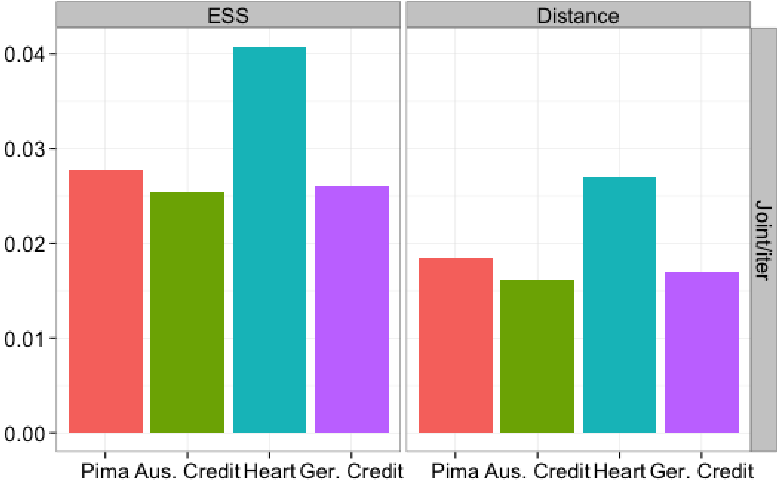
- **Pima Indian data** ($n = 532, p = 8$): outcome is diabetes; covariates include BMI, age, number of pregnancies
- **Australian credit data** ($n = 690, p = 14$): outcome is credit approval; 14 generic covariates
- **Heart disease data** ($n = 270, p = 13$): outcome is heart disease; covariates include 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

Probit performance: median values in 10 runs



Probit performance: relative

Standardize for run time: $\frac{\text{ESS/second, joint}}{\text{ESS/second, iterative}}$ and $\frac{\text{Dist./second, joint}}{\text{Dist./second, iterative}}$



From probit to logit

Extend auxiliary variables to logistic regression with another level to model differing variances of the error terms:

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

Equivalent to logit model with $y_i \sim \text{Bernoulli}\,(p_i = \text{expit}(\mathbf{x}_i\boldsymbol{\beta}))$ because ϵ_i has a logistic distribution (Andrews & Mallows, 1974) and CDF of logistic is expit function:

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

Logistic Gibbs

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

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

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

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

(3) $\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{y}) \propto p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda})p(\boldsymbol{\lambda}) \cong$ ind. normal $\times$ 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")

Logistic sampler comparison

Iterative updates (not analyzed in paper)**:**

- Iterate block Gibbs updates $\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda}, \mathbf{y})$, $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$, $\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{y})$
- Variance of $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$ is λ_i with expected value $\pi^2/3$

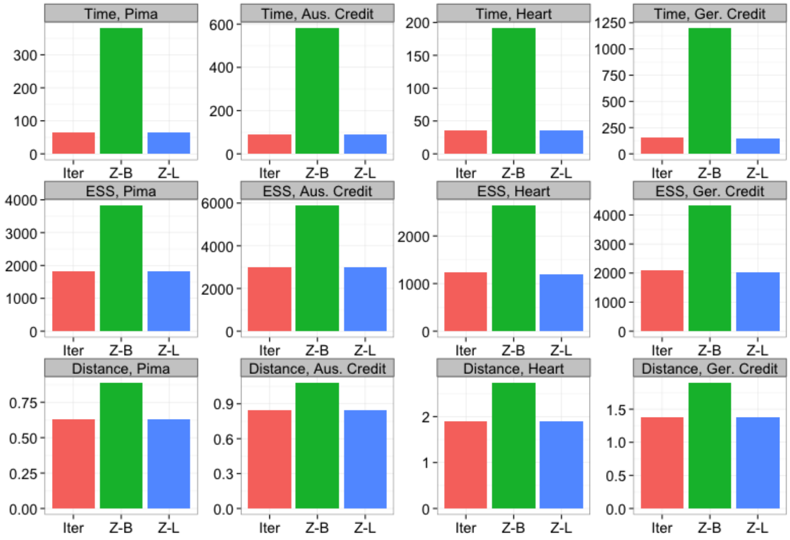
Joint updating scheme $(\mathbf{z}, \boldsymbol{\lambda})$**:**

- Iterate block Gibbs updates $\pi(\mathbf{z}, \boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})}_{\text{trunc ind logistic}} \underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z})}_{\text{normal}\times\text{KS}^2}$, then $\underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda})}_{\text{normal}}$
- Variance of $\pi(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})$ is $\pi^2/3$ – little gain by marginalizing?

Joint updating scheme $(\mathbf{z}, \boldsymbol{\beta})$**:**

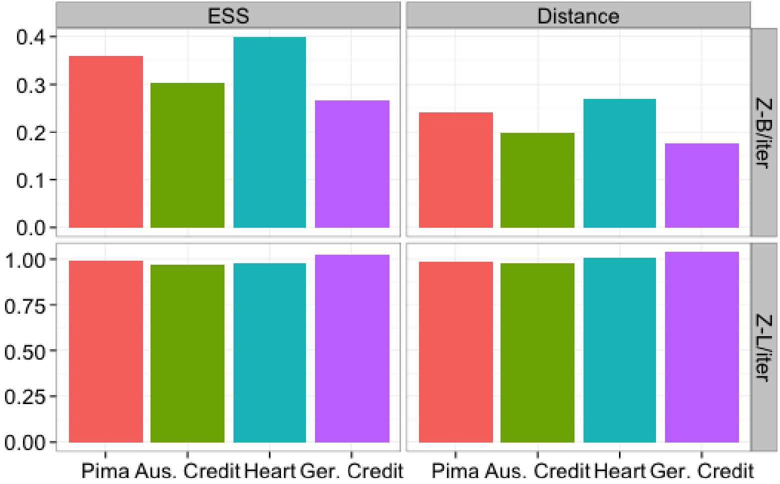
- Iterate Gibbs updates $\pi(\mathbf{z}, \boldsymbol{\beta} \mid \boldsymbol{\lambda}, \mathbf{y}) = \underbrace{\pi(\mathbf{z} \mid \boldsymbol{\lambda}, \mathbf{y})}_{\text{trunc mv normal}} \underbrace{\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\lambda})}_{\text{normal}}$, then $\underbrace{\pi(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \mathbf{z})}_{\text{normal}\times\text{KS}^2}$
- Note that $\pi(\mathbf{z} \mid \boldsymbol{\lambda}, \mathbf{y})$ will require Gibbsing through $\pi(z_i \mid \mathbf{z}_{-i}, \boldsymbol{\lambda}, \mathbf{y})$, but variance is $> \lambda_i$

Logit performance: median values in 10 runs



Logit performance: relative

Standardize for run time: $\frac{\text{ESS/second, joint}}{\text{ESS/second, iterative}}$ and $\frac{\text{Dist./second, joint}}{\text{Dist./second, iterative}}$



Concluding thoughts

- Latent variables can induce convenient conditional distributions to make MCMC sampling tractable for Bayesian models of binary data
  - In these cases, all conditionals can be sampled from without Metropolis-Hastings
- Joint updating to increase variance in Gibbs sampling might make sense theoretically. . .
  - . . . but only the scheme updating $(\mathbf{z}, \boldsymbol{\lambda})$ jointly in logistic regression was competitive with blocked iterative updates
  - Don't replace independent truncated univariate distributions with a truncated multivariate normal!
- Auxiliary variable technique H&H introduced for logistic regression extends straightforwardly to Bayesian model uncertainty situations, polytomous outcomes