Session 4: Auxiliary Variables / Data Augmentation

Bayesian Inference for the Social Sciences
November 3, 2017

Kevin Quinn

Auxiliary Variables / Data Augmentation

In some situations it becomes useful to include additional random variables in the sampling scheme even though these additional variables are not of direct interest themselves

In the MCMC literature this is known as using auxiliary variables

A particular form of this is known as data augmentation

Using auxiliary variables may:

- make full conditional distributions take known parametric forms
- improve convergence
- allow you to deal with missing data

Auxiliary Variables / Data Augmentation

Basic idea behind introducing auxiliary variables:

Auxiliary Variables / Data Augmentation

Basic idea behind introducing auxiliary variables:

Suppose we would like to sample from a (possibly multivariate) density p

Auxiliary Variables / Data Augmentation

Basic idea behind introducing auxiliary variables:

Suppose we would like to sample from a (possibly multivariate) density p

Instead of working directly with p we are going to work with a density f that satisfies:

$$\int_Z f(\theta, z) dz = p(\theta)$$

Auxiliary Variables / Data Augmentation

Basic idea behind introducing auxiliary variables:

Suppose we would like to sample from a (possibly multivariate) density p

Instead of working directly with p we are going to work with a density f that satisfies:

$$\int_Z f(\theta, z)dz = p(\theta)$$

f is called a completion of p and is chosen so that its full conditional distributions are easy to sample from

Auxiliary Variables / Data Augmentation

Basic idea behind introducing auxiliary variables:

Suppose we would like to sample from a (possibly multivariate) density p

Instead of working directly with p we are going to work with a density f that satisfies:

$$\int_Z f(\theta, z) dz = p(\theta)$$

f is called a completion of p and is chosen so that its full conditional distributions are easy to sample from

Since f is a completion of p and f's full conditionals are easy to sample from we can get a sample of (θ, z) from f using the Gibbs sampling algorithm and then ignore the z draws to get a sample of θ from p

Auxiliary Variables / Data Augmentation

Slice sampling provides a nice illustration of how this works.

Suppose we want to sample from

$$p(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right)$$

Auxiliary Variables / Data Augmentation

Slice sampling provides a nice illustration of how this works.

Suppose we want to sample from

$$p(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right)$$

p is a standard normal density but we'll ignore that and instead work with the following completion of p

$$f(\theta, z) \propto \mathbb{I}\left[0 \leq z \leq \exp\left(-\frac{\theta^2}{2}\right)\right]$$

Auxiliary Variables / Data Augmentation

Slice sampling provides a nice illustration of how this works.

Suppose we want to sample from

$$p(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right)$$

p is a standard normal density but we'll ignore that and instead work with the following completion of p

$$f(\theta, z) \propto \mathbb{I}\left[0 \leq z \leq \exp\left(-\frac{\theta^2}{2}\right)\right]$$

To verify that f is a completion of p note that

$$\begin{aligned}\int_{-\infty}^{\infty} f(\theta, z)dz &\propto \int_{-\infty}^{\infty} \mathbb{I}\left[0 \leq z \leq \exp\left(-\frac{\theta^2}{2}\right)\right] dz \\ &\propto \int_{0}^{\exp(-\theta^2/2)} 1 dz \\ &\propto \exp\left[-\theta^2/2\right]\end{aligned}$$

which is propotional to p

Auxiliary Variables / Data Augmentation

The full conditionals also take particularly nice forms:

$$[Z|\theta] \sim \mathcal{U}nif\left(0, \exp\left[-\theta^2/2\right]\right)$$

$$[\theta|z] \sim \mathcal{U}nif\left(-\sqrt{-2\log(z)}, \sqrt{-2\log(z)}\right)$$

Auxiliary Variables / Data Augmentation

The full conditionals also take particularly nice forms:

$$[Z|\theta] \sim \mathcal{U}nif\left(0, \exp\left[-\theta^2/2\right]\right)$$

$$[\theta|z] \sim \mathcal{U}nif\left(-\sqrt{-2\log(z)}, \sqrt{-2\log(z)}\right)$$

Punchline: It's possible to sample from a normal distribution using a Gibbs sampler based on Uniform full conditionals!!!

The Probit Model

The probit model is:

$$y_i \stackrel{ind.}{\sim} \mathcal{B}ernoulli(\pi_i) \quad i = 1, \ldots, n$$

$$\pi_i = \Phi(\mathbf{x}_i' \beta) \quad i = 1, \ldots, n$$

The Probit Model

The probit model is:

$$y_i \overset{ind.}{\sim} \mathcal{B}ernoulli(\pi_i) \quad i = 1, \ldots, n$$

$$\pi_i = \Phi(\mathbf{x}_i'\beta) \quad i = 1, \ldots, n$$

The probit sampling density is:

$$p(\mathbf{y}|\beta) = \prod_{i=1}^{n} \Phi(\mathbf{x}_i'\beta)^{y_i}[1 - \Phi(\mathbf{x}_i'\beta)]^{(1-y_i)}$$

The Probit Model

The probit model is:

$$y_i \overset{ind.}{\sim} \mathcal{B}ernoulli(\pi_i) \quad i = 1, \ldots, n$$

$$\pi_i = \Phi(\mathbf{x}_i' \beta) \quad i = 1, \ldots, n$$

The probit sampling density is:

$$p(\mathbf{y}|\beta) = \prod_{i=1}^{n} \Phi(\mathbf{x}_i'\beta)^{y_i}[1 - \Phi(\mathbf{x}_i'\beta)]^{(1-y_i)}$$

Regardless of the prior chosen for β the resulting posterior $p(\beta|\mathbf{y})$ does *not* have nice full conditional distributions.

The Probit Model

Note that the probit sampling density contains the standard normal cdf Φ

The Probit Model

Note that the probit sampling density contains the standard normal cdf Φ– This suggests a potential completion

The Probit Model

Note that the probit sampling density contains the standard normal cdf Φ– This suggests a potential completion

Instead of working with $p(\beta|\mathbf{y})$ let's work with (Albert and Chib, 1993):

$$f(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y}) \propto \left\{ \prod_{i=1}^{n} \left[\mathbb{I}(z_i > 0)\mathbb{I}(y_i = 1) + \mathbb{I}(z_i \leq 0)\mathbb{I}(y_i = 0) \right] \phi(z_i|\mathbf{x}_i'\boldsymbol{\beta}, 1) \right\} p(\boldsymbol{\beta})$$

where $\phi(\cdot|\mu, \sigma^2)$ is a normal density with mean μ and variance σ^2

The Probit Model

Note that the probit sampling density contains the standard normal cdf Φ– This suggests a potential completion

Instead of working with $p(\beta|\mathbf{y})$ let's work with (Albert and Chib, 1993):

$$f(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y}) \propto \left\{ \prod_{i=1}^{n} \left[\mathbb{I}(z_i > 0)\mathbb{I}(y_i = 1) + \mathbb{I}(z_i \leq 0)\mathbb{I}(y_i = 0) \right] \phi(z_i|\mathbf{x}_i'\boldsymbol{\beta}, 1) \right\} p(\boldsymbol{\beta})$$

where $\phi(\cdot|\mu, \sigma^2)$ is a normal density with mean μ and variance σ^2

In what follows we'll assume that $p(\boldsymbol{\beta})$ is a normal density with mean $\mathbf{m}$ and variance-covariance matrix $\mathbf{V}$

The Probit Model

First of all, is $f(\beta, \mathbf{z}|\mathbf{y})$ a completion of $p(\beta|\mathbf{y})$?

The Probit Model

First of all, is $f(\beta, \mathbf{z}|\mathbf{y})$ a completion of $p(\beta|\mathbf{y})$?

- Yes– straightforward to see that $\int_Z f(\beta, \mathbf{z}|\mathbf{y})dz = p(\beta|\mathbf{y})$

The Probit Model

First of all, is $f(\beta, \mathbf{z}|\mathbf{y})$ a completion of $p(\beta|\mathbf{y})$?

- Yes– straightforward to see that $\int_Z f(\beta, \mathbf{z}|\mathbf{y}) dz = p(\beta|\mathbf{y})$

Are the full conditionals easy to sample from?

The Probit Model

First of all, is $f(\beta, \mathbf{z}|\mathbf{y})$ a completion of $p(\beta|\mathbf{y})$?

- Yes– straightforward to see that $\int_Z f(\beta, \mathbf{z}|\mathbf{y})dz = p(\beta|\mathbf{y})$

Are the full conditionals easy to sample from?

- Yes–

$$z_i|\beta, \mathbf{y} \sim \begin{cases} \mathcal{TN}_{(>0)}(\mathbf{x}_i'\beta, 1) & \text{if } y_i = 1 \\ \mathcal{TN}_{(\leq 0)}(\mathbf{x}_i'\beta, 1) & \text{if } y_i = 0 \end{cases}$$

The Probit Model

First of all, is $f(\beta, \mathbf{z}|\mathbf{y})$ a completion of $p(\beta|\mathbf{y})$?

- Yes– straightforward to see that $\int_Z f(\beta, \mathbf{z}|\mathbf{y})dz = p(\beta|\mathbf{y})$

Are the full conditionals easy to sample from?

- Yes–

$$z_i|\beta, \mathbf{y} \sim \begin{cases} \mathcal{TN}_{(>0)}(\mathbf{x}_i'\beta, 1) & \text{if } y_i = 1 \\ \mathcal{TN}_{(\leq 0)}(\mathbf{x}_i'\beta, 1) & \text{if } y_i = 0 \end{cases}$$

$$\beta|\mathbf{z}, \mathbf{y} \sim \mathcal{N}(\tilde{\beta}, \mathbf{\Sigma}_\beta)$$

where $\mathbf{\Sigma}_\beta = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$ and $\tilde{\beta} = \mathbf{\Sigma}_\beta(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{z})$

The Probit Model

What's the intuition here?

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the latent variable $\mathbf{z}$ satisfies a regression relationship:

$$z_i = \mathbf{x}_i'\beta + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1) \quad i = 1, \ldots, n$$

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the latent variable $\mathbf{z}$ satisfies a regression relationship:

$$z_i = \mathbf{x}_i' \beta + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1) \quad i = 1, \ldots, n$$

If we could observed the zs then we would have a simple linear regression model

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the latent variable $\mathbf{z}$ satisfies a regression relationship:

$$z_i = \mathbf{x}_i'\beta + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1) \quad i = 1, \ldots, n$$

If we could observed the zs then we would have a simple linear regression model

While we can't directly observe the zs we do know how they are distributed given $\mathbf{x}_i'$ and β (and the modeling assumptions)

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the latent variable $\mathbf{z}$ satisfies a regression relationship:

$$z_i = \mathbf{x}_i'\beta + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1) \quad i = 1, \ldots, n$$

If we could observed the zs then we would have a simple linear regression model

While we can't directly observe the zs we do know how they are distributed given $\mathbf{x}_i'$ and β (and the modeling assumptions)

The auxiliary variable scheme samples $\mathbf{z}$ from its full conditional

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the latent variable $\mathbf{z}$ satisfies a regression relationship:

$$z_i = \mathbf{x}_i'\beta + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1) \quad i = 1, \ldots, n$$

If we could observed the zs then we would have a simple linear regression model

While we can't directly observe the zs we do know how they are distributed given $\mathbf{x}_i'$ and β (and the modeling assumptions)

The auxiliary variable scheme samples $\mathbf{z}$ from its full conditional

Given $\mathbf{z}$, the full conditional for β is the full conditional from a linear regression

The Probit Model

What's the intuition here?

Recall that the probit model can be motivated as a latent variable model:

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the latent variable $\mathbf{z}$ satisfies a regression relationship:

$$z_i = \mathbf{x}_i'\beta + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1) \quad i = 1, \ldots, n$$

If we could observed the zs then we would have a simple linear regression model

While we can't directly observe the zs we do know how they are distributed given $\mathbf{x}_i'$ and β (and the modeling assumptions)

The auxiliary variable scheme samples $\mathbf{z}$ from its full conditional

Given $\mathbf{z}$, the full conditional for β is the full conditional from a linear regression

Averaging over the draws of the $\mathbf{z}$s gives the same answer as would be arrived at by working directly with $p(\beta|\mathbf{y})$

Estimating Ideal Points of U.S. Supreme Court Justices

- $X \subset \mathbb{R}^D$ the policy space
- $\boldsymbol{\theta}_j \in X$ justice j's ideal point
- $\mathbf{x}_k^{(a)} \in X$ policy outcome if Court affirms the lower court's ruling on case k
- $\mathbf{x}_k^{(r)} \in X$ policy outcome if Court reverses the lower court's ruling on case k
- $k = 1, 2, \ldots, K$ (cases)
- $j = 1, 2, \ldots, J$ (justices)

Estimating Ideal Points of U.S. Supreme Court Justices

- (Random) utility of voting to affirm:

$$u_{k,j}^{(a)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)}$$

- (Random) utility of voting to reverse:

$$u_{k,j}^{(r)} = -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 + \delta_{k,j}^{(r)}$$

- Difference in utility:

$$\begin{aligned}
z_{k,j} &= u_{k,j}^{(a)} - u_{k,j}^{(r)} \\
&= -\|\boldsymbol{\theta}_j - \mathbf{x}_k^{(a)}\|^2 + \delta_{k,j}^{(a)} + \|\boldsymbol{\theta}_j - \mathbf{x}_k^{(r)}\|^2 - \delta_{k,j}^{(r)} \\
&= \left[\mathbf{x}_k^{\prime(r)}\mathbf{x}_k^{(r)} - \mathbf{x}_k^{\prime(a)}\mathbf{x}_k^{(a)}\right] + 2\boldsymbol{\theta}_j^{\prime}\left[\mathbf{x}_k^{(a)} - \mathbf{x}_k^{(r)}\right] + \left[\delta_{k,j}^{(a)} - \delta_{k,j}^{(r)}\right] \\
&= \alpha_k + \boldsymbol{\beta}_k^{\prime}\boldsymbol{\theta}_j + \varepsilon_{k,j} \qquad \varepsilon_{k,j} \overset{iid}{\sim} \mathcal{N}(0, 1)
\end{aligned}$$

Estimating Ideal Points of U.S. Supreme Court Justices

- Observation equation:

$$y_{k,j} = \begin{cases} 1 \text{ (affirm)} & \text{if} \quad z_{k,j} > 0 \\ 0 \text{ (reverse)} & \text{if} \quad z_{k,j} \leq 0 \end{cases}$$

- Distributional assumption:

$$y_{k,j} | \alpha_k, \boldsymbol{\beta}_k, \boldsymbol{\theta}_j \overset{ind.}{\sim} \mathcal{B}ernoulli(\pi_{k,j})$$

where

$$\pi_{k,j} = \Phi(\alpha_k + \boldsymbol{\beta}_k' \boldsymbol{\theta}_j)$$

- Sampling density:

$$p(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{j=1}^{J} \Phi(\alpha_k + \boldsymbol{\beta}_k' \boldsymbol{\theta}_j)^{y_{k,j}} \times \left[1 - \Phi(\alpha_k + \boldsymbol{\beta}_k' \boldsymbol{\theta}_j)\right]^{1-y_{k,j}}$$

Estimating Ideal Points of U.S. Supreme Court Justices

Standard priors:

- Ideal Points

$$\boldsymbol{\theta}_j \overset{iid}{\sim} \mathcal{N}(\mathbf{t}_0, \mathbf{T}_0) \quad j = 1, 2, \ldots, J$$

- Case Parameters

$$\boldsymbol{\eta}_k = \begin{bmatrix} \alpha_k \\ \boldsymbol{\beta}_k \end{bmatrix} \overset{iid}{\sim} \mathcal{N}_{D+1}(\mathbf{b}_0, \mathbf{B}_0) \quad k = 1, 2, \ldots, K$$

Estimating Ideal Points of U.S. Supreme Court Justices

Model Fitting::

1.
$$[z_{k,j}|\boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{Y}] \sim \begin{cases} \mathcal{TN}_{(>0)}(\alpha_k + \boldsymbol{\beta}_k' \boldsymbol{\theta}_j,\ 1) & \text{if } y_i = 1 \\ \mathcal{TN}_{(\leq 0)}(\alpha_k + \boldsymbol{\beta}_k' \boldsymbol{\theta}_j,\ 1) & \text{if } y_i = 0 \end{cases}$$
$$k = 1, 2, \ldots, K \quad j = 1, 2, \ldots, J$$

2.
$$[\boldsymbol{\eta}_k|\boldsymbol{\theta}, \mathbf{Z}, \mathbf{Y}] \sim \mathcal{N}(\mathbf{e}, \mathbf{E}), \quad k = 1, 2, \ldots, K$$

where $\mathbf{e} = \mathbf{E}\left[\boldsymbol{\theta}^{*'}\mathbf{z}_{k,\cdot} + \mathbf{B}_0^{-1}\mathbf{b}_0\right]$, $\mathbf{E} = \left[\boldsymbol{\theta}^{*'}\boldsymbol{\theta}^* + \mathbf{B}_0^{-1}\right]^{-1}$ and $\boldsymbol{\theta}^*$ is the ideal point matrix with 1s appended in the first column.

3.
$$[\boldsymbol{\theta}_j|\boldsymbol{\eta}, \mathbf{Z}, \mathbf{Y}] \sim \mathcal{N}(\mathbf{t}, \mathbf{T}), \quad j = 1, \ldots, J$$

where $\mathbf{t} = \mathbf{T}\left[\boldsymbol{\beta}'(\boldsymbol{\alpha} - \mathbf{z}_{\cdot,j}') + \mathbf{T}_0^{-1}\mathbf{t}_0\right]$, $\mathbf{T} = [\boldsymbol{\beta}'\boldsymbol{\beta} + \mathbf{T}_0]^{-1}$.

Estimating Ideal Points of U.S. Supreme Court Justices

`R` example

Missing Bivariate Normal Data

`R` example