

# Bayesian Bayesics

Justin Tobias

Camp Resources Tutorial

August 7, 2012

## Outline

- 1 Introduction
- 2 Preliminaries and Bayes Theorem
- 3 The Gibbs Algorithm
  - The Gibbs Kernel
  - Example: Simple Bivariate Normal Sampling
- 4 Gibbs Sampling in a Linear Regression model
  - Illustrative Application with Wage Data
    - Prediction via Composition
- 5 Binary Choice Model
- 6 Ordinal and Multinomial Choice
- 7 Endogenous Dummy Variable, Continuous Outcome Model
  - Extensions and Recent Work
    - Flexible Bayes
    - Imperfect Instruments
- 8 Conclusion

## Introduction

- Among those of us with sufficient experience to have witnessed trends in the profession (read: old folks like me), most would probably agree that Bayesian methods have certainly become more popular over the last 15-20 years.
- When presenting my work straight out of graduate school, I found that many of those who had not yet fallen asleep proved to be totally unfamiliar with Bayes - even Bayes theorem (!) - and were very skeptical of my results (and me).

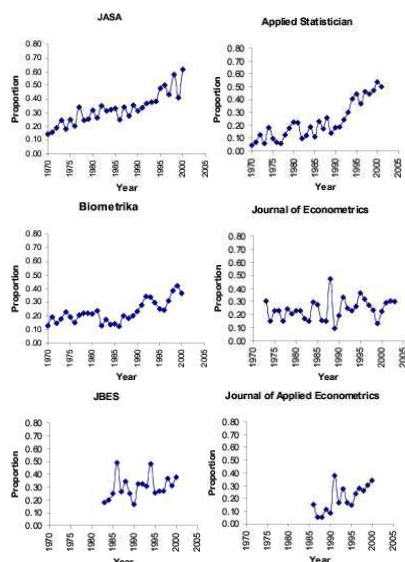
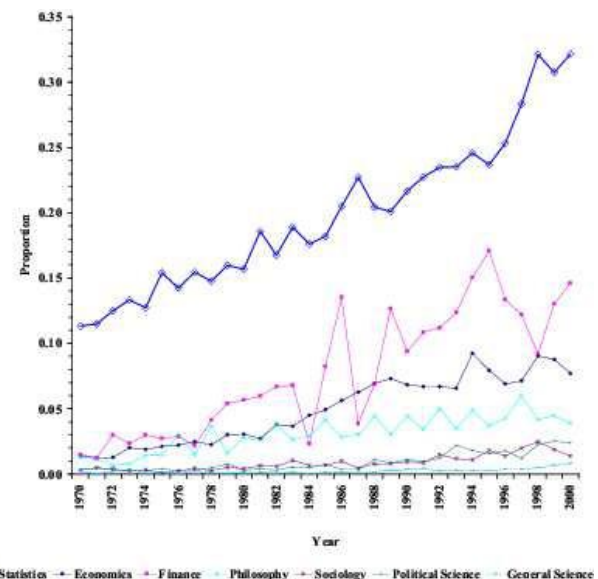
## Introduction

- Now, I find that more are familiar with the basics of Bayes, yet most continue to have an inherent distrust of priors (and me).
- The perceived growth in Bayes is certainly true within Statistics, also true among Economics and Econometrics generally, and my priors (coupled with a few data points) suggest that is is also true, though perhaps to a lesser extent, within Environmental / Resource Economics.


- Poirier (2006) offers some evidence that this has indeed been the case in many fields.


 Poirier, D.J. (2006). "The Growth of Bayesian Methods in Statistics and Economics Since 1970," *Bayesian Analysis*, 969-980.

- He searches for the appearance of "Bayes" or "Bayesian" in articles and records the fraction of such articles by field and by journal over time:
  - **Statistics** (23 journals)
  - **Finance** (4 Journals)
  - **Economics** (41 Journals)
  - **Philosophy** (24 Journals)
  - Political Science (40 Journals)
  - Sociology (39 Journals)
  - General Science (7 Journals)
- The following figures are all taken from the Poirier article:



- So, the material that I am about to review appears to be growing in popularity, but a relevant question is *why*?
- Within Economics and other fields, a key reason for the expansion of Bayesian work is the development of Markov Chain Monte Carlo (MCMC) methods.
- Two key players here are the **Gibbs Sampler** and the **Metropolis-Hastings** algorithm.

 Casella, G. and E. George (1992). "Explaining the Gibbs Sampler." *The American Statistician*, 167-174.

 Chib, S and E. Greenberg (1995). "Understanding the Metropolis-Hastings Algorithm." *The American Statistician*, 327-335.

Below are several other papers coincident with the start of the “MCMC Revolution:”

- 📄 Gelfand, A.E. and A.F.M. Smith (1990). “Sampling-Based Approaches to Calculating Marginal Densities” *Journal of the American Statistical Association*, 398-509.
- 📄 Gelfand, A.E., S.E. Hills, A. Racine-Poon and A.F.M. Smith (1990). ‘Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling’. *Journal of the American Statistical Association*, 412: 972-85.
- 📄 Chib, S. (1992), ‘Bayes Inference in the Tobit Censored Regression Model’. *Journal of Econometrics*, 51: 79-99.
- 📄 Albert, J. and S. Chib (1993). ‘Bayesian Analysis of Binary and Polychotomous Response Data’. *Journal of the American Statistical Association*, 88: 669-79.

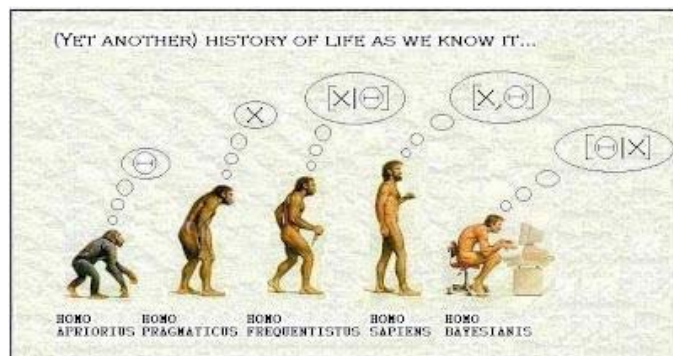
## Review of Basic Framework

- Quantities to become known under sampling are denoted by the  $n$ -dimensional vector  $y$ .
- The remaining unknown quantities are denoted by the  $K$ -dimensional vector  $\theta \in \Theta \subseteq \mathcal{R}^K$ .
- Consider the joint density of observables  $y$  and unobservables  $\theta$ :

$$p(y, \theta)$$

## Review of Basic Framework

(Not sure to whom credit should be given for this, but it is not my own creation:)



## Review of Basic Framework

- Standard manipulations show:

$$p(y, \theta) = p(\theta)p(y|\theta) = p(y)p(\theta|y),$$

where

- $p(\theta)$  is the **prior density**
- $p(\theta|y)$  is the **posterior density**
- $p(y|\theta)$  is the **likelihood function**. [Viewed as a function of  $\theta$ , we write this as  $L(\theta)$ ].

## Review of Basic Framework

We also note

$$\begin{aligned} p(y) &= \int_{\Theta} p(\theta)p(y|\theta)d\theta \\ &= \int_{\Theta} p(\theta)L(\theta)d\theta \end{aligned}$$

is the **marginal density of the observed data**.

## Bayes Theorem

**Bayes' theorem for densities** follows immediately:

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)L(\theta)}{p(y)} \\ &\propto p(\theta)L(\theta). \end{aligned}$$

- We focus on the posterior **up to proportionality** ( $\propto$ ) on the right-hand side.
- Often, the kernel of the posterior will take on a familiar form, whence the normalizing constant of the posterior can be deduced.
- The shape of the posterior can be learned by plotting the right hand side of this expression when  $k = 1$  or  $k = 2$ .
- In these cases, the normalizing constant  $[p(y)]^{-1}$  can be obtained numerically (e.g., a trapezoidal rule or Simpson's rule).

## Bayes Theorem

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

- In most non-trivial situations, the integration required to calculate posterior statistics of interest **cannot be performed analytically**.
- Enter MCMC: To calculate the desired quantities, we generate a series of simulations which converge in distribution to the joint posterior  $p(\theta|y)$ .
- We can then use the post-convergence simulations to calculate posterior means, standard deviations, quantiles and entire marginal posterior distributions.

## The Gibbs Algorithm

## The Gibbs Algorithm

- As analytic results are seldom available in problems of even moderate complexity, we commonly rely on numerical strategies to generate simulations from the joint posterior  $p(\theta|y)$ . These simulations can then be used to calculate the desired posterior features.
- One such device is the **Gibbs Sampler**.

Let  $\theta$  be a  $K \times 1$  parameter vector with associated posterior distribution  $p(\theta|y)$  and write


$$\theta = [\theta^1 \ \theta^2 \ \dots \ \theta^K].$$


(We use superscripts to denote elements of the parameter vector and will employ subscripts to denote iterations in the algorithm.)

The **Gibbs sampling algorithm** proceeds as follows:

- (i) Select an initial parameter vector  $\theta_0 = [\theta_0^1 \ \theta_0^2 \ \dots \ \theta_0^K]$ . This initial condition could be arbitrarily chosen, sampled from the prior, or perhaps could be obtained from a crude estimation method such as least-squares.
  - (1) Sample  $\theta_1^1$  from the **complete posterior conditional** density:  $p(\theta^1 | \theta^2 = \theta_0^2, \theta^3 = \theta_0^3, \dots, \theta^K = \theta_0^K, y)$ .
  - (2) Sample  $\theta_1^2$  from  $p(\theta^2 | \theta^1 = \theta_1^1, \theta^3 = \theta_0^3, \dots, \theta^K = \theta_0^K, y)$
  - $\vdots$
  - (K) Sample  $\theta_1^K$  from  $p(\theta^K | \theta^1 = \theta_1^1, \theta^2 = \theta_1^2, \dots, \theta^{K-1} = \theta_1^{K-1}, y)$
- (ii) Repeatedly cycle through (1)  $\rightarrow$  (K) to obtain  $\theta_2 = [\theta_2^1 \ \theta_2^2 \ \dots \ \theta_2^K]$ ,  $\theta_3$ , etc., **always conditioning on the most recent values of the parameters drawn** [e.g., to obtain  $\theta_2^1$ , draw from  $p(\theta^1 | \theta^2 = \theta_1^2, \theta^3 = \theta_1^3, \dots, \theta^K = \theta_1^K, y)$ , etc.].

- Under standard regularity conditions, the sequence of draws produced from this algorithm will act as (correlated) draws from  $p(\theta|y)$ .

 Tierney, L. (1994). "Markov Chains for Exploring Posterior Distributions" (with discussion and rejoinder). *Annals of Statistics*, 1701-1762.

 Roberts, G.O. and A.F.M. Smith (1994). "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms" *Stochastic Processes and their Applications*, 207-216.

 Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics* (section 4.5 in particular).

- To implement the Gibbs sampler we require the ability to draw from the posterior conditionals of the model.
- Although the joint posterior density  $p(\theta|y)$  may often be intractable, the complete conditionals  $\{p(\theta^j | \theta^{-j}, y)\}_{j=1}^K$ , (with  $\theta^{-j}$  denoting all parameters other than  $\theta^j$ ) prove to be of standard forms in many cases!

- MCMC algorithms construct a **transition kernel**  $K$  to ensure that the joint posterior is a stationary distribution of the Markov chain.
- The **Gibbs Sampling** algorithm constructs this transition kernel by *sampling from the conditionals of the target (posterior) distribution*.
- To provide a specific example, consider a bivariate distribution  $p(y_1, y_2)$ .
- Further, apply the transition kernel

$$K(x_1, x_2, y_1, y_2) = p_{1|2}^*(y_1|x_2)p_{2|1}^*(y_2|y_1).$$

- That is, if you are currently at  $(x_1, x_2)$ , then the probability that you will be at  $(y_1, y_2)$  can be surmised from the conditional distributions of  $p$ ,  $p_{1|2}^*(Y_1|Y_2 = x_2)$  and  $p_{2|1}^*(Y_2|y_1)$  (where  $y_1$  refers to the value realized from the first step).

It is reasonably straightforward to show that the target distribution  $p(y_1, y_2)$  is a stationary distribution under this transition kernel:

To this end, note

$$\begin{aligned} \int K(x, y)p(x)dx &= \int \int p_{1|2}^*(y_1|x_2)p_{2|1}^*(y_2|y_1)p(x_1, x_2)dx_1dx_2 \\ &= \int \int p_{1|2}^*(y_1|x_2)p_{2|1}^*(y_2|y_1)p_{1|2}^*(x_1|x_2)p_2^*(x_2)dx_1dx_2 \\ &= \int p_{1|2}^*(y_1|x_2)p_{2|1}^*(y_2|y_1)p_2^*(x_2)dx_2 \\ &= p_{2|1}^*(y_2|y_1) \int p_{1|2}^*(y_1|x_2)p_2^*(x_2)dx_2 \\ &= p_{2|1}^*(y_2|y_1)p_1^*(y_1) \\ &= p(y_1, y_2). \end{aligned}$$

## Simple Bivariate Normal Sampling

- We now turn to, perhaps, the simplest example of the Gibbs sampler, and illustrate how the algorithm is implemented within the context of this model.
- We suppose that some problem of interest generates a posterior distribution of the form:

$$p(\theta_1, \theta_2|y) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

where  $\rho$  is *known*.

- We will illustrate how the Gibbs sampler can be employed to fit this model (even though sampling from a bivariate normal is a trivial thing to do!)

## Simple Bivariate Normal Sampling

- To begin, we must set a starting value for *either*  $\theta_1$  or  $\theta_2$ .
- It doesn't matter which we choose - the algorithm will work either way. So, let's say that we set  $\theta_2 = c$  to start.
- To implement the Gibbs sampler, we must derive the conditional posterior distributions  $p(\theta_1|\theta_2, y)$  and  $p(\theta_2|\theta_1, y)$ . These are readily available using properties of the multivariate normal distribution:

$$\theta_1|\theta_2, y \sim N(\rho\theta_2, 1 - \rho^2),$$

and

$$\theta_2|\theta_1, y \sim N(\rho\theta_1, 1 - \rho^2).$$

## Simple Bivariate Normal Sampling

So, the **first** iteration of the Gibbs sampler will proceed as follows:

1 Set  $\theta_2 = c$ .

2 Sample

$$\theta_1^* \sim N(\rho c, 1 - \rho^2).$$

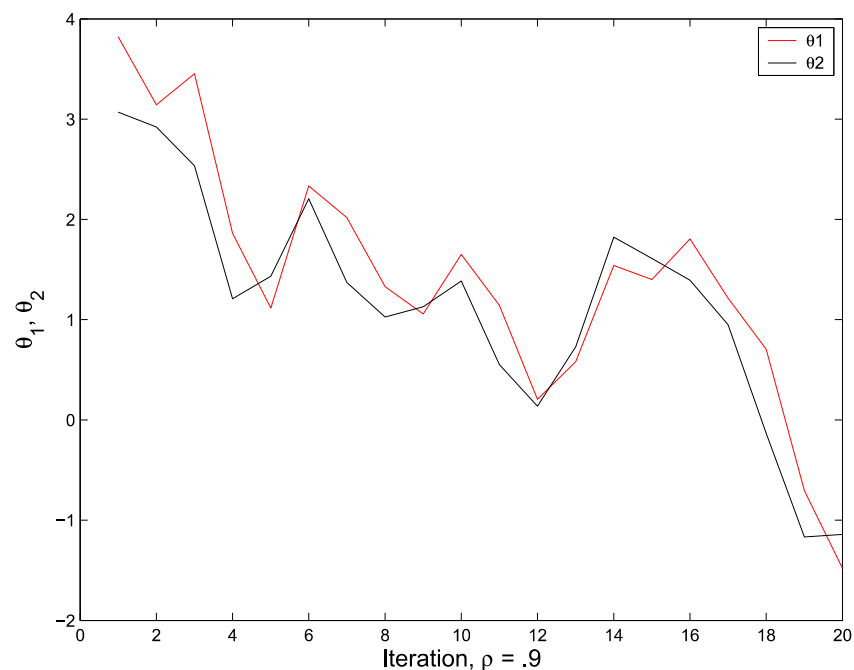
3 Sample

$$\theta_2^* \sim N(\rho \theta_1^*, 1 - \rho^2).$$

Thus,  $(\theta_1^*, \theta_2^*)$  denotes our Gibbs sample after the first iteration.

- One needs to get rid of some of the initial simulations and use the latter set of simulations to calculate quantities of interest.
- Remember that this is an **iterative** algorithm - we must first converge to the target (posterior) distribution, and once we have arrived at this target distribution, the subsequent draws will be draws from  $p(\theta|y)$ .
- This “pre-convergence” period is called the **burn-in**, and the burn-in draws should be discarded.
- A sketch of a MATLAB program that does all of these things is provided on the following page:

```
rho = c;
iter = 1000;
burn = 100;
theta1keep = zeros(iter-burn,1);
theta2keep = theta1keep;
theta2draw = 4;
for i=1:iter;
    theta1draw = rho*theta2draw + sqrt(1-rho^2)*randn(1,1);
    theta2draw = rho*theta1draw + sqrt(1-rho^2)*randn(1,1);
    if i > burn;
        theta1keep(i-burn) = theta1draw;
        theta2keep(i-burn) = theta2draw;
    end;
end;
```



# Gibbs in the Linear Regression Model

## Gibbs Sampling in the Linear Regression Model

Consider the regression model (where  $y$  is  $n \times 1$ )

$$y|X, \beta, \Sigma \sim N(X\beta, \Sigma)$$

under the proper prior

$$\beta \sim N(\mu_\beta, V_\beta), \quad p(\Sigma)$$

One can show (completing the square on  $\beta$  plus a bit of algebra):

$$\beta|\Sigma, y \sim N(D_\beta d_\beta, D_\beta)$$

where

$$D_\beta = \left( X' \Sigma^{-1} X + V_\beta^{-1} \right)^{-1}, \quad d_\beta = X' \Sigma^{-1} y + V_\beta^{-1} \mu_\beta.$$



Lindley, D.V. and A.F.M. Smith (1972). "Bayes Estimates for the Linear Model" *JRSS B* 1-41.

## Gibbs Sampling in the Linear Regression Model

Consider the regression model

$$y|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n)$$

under the priors

$$\beta \sim N(\mu_\beta, V_\beta), \quad \sigma^2 \sim IG(a, b).$$

- We seek to show how Gibbs sampling can be used to calculate features of the joint posterior distribution.

- To implement the sampler, we need to derive two things:

1

$$p(\beta|\sigma^2, y).$$

2

$$p(\sigma^2|\beta, y).$$

The first of these can be obtained by applying our previous theorem (with  $\Sigma = \sigma^2 I_n$ ). Specifically, we obtain:

$$\beta|\sigma^2, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = \left( X' X / \sigma^2 + V_\beta^{-1} \right)^{-1}, \quad d_\beta = X' y / \sigma^2 + V_\beta^{-1} \mu_\beta.$$

As for the posterior conditional for  $\sigma^2$ , note

$$p(\beta, \sigma^2 | y) \propto p(\beta) p(\sigma^2) p(y | \beta, \sigma^2).$$

Since  $p(\sigma^2 | \beta, y)$  is proportional to the joint posterior above, it follows that

$$\begin{aligned} p(\sigma^2 | \beta, y) &\propto p(\sigma^2) p(y | \beta, \sigma^2) \\ &\propto [\sigma^2]^{-(a+1)} \exp\left(-\frac{1}{b\sigma^2}\right) \\ &\quad \times [\sigma^2]^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \\ &= [\sigma^2]^{-([n/2]+a+1)} \\ &\quad \times \exp\left(-\frac{1}{\sigma^2} \left[b^{-1} + \frac{1}{2}(y - X\beta)'(y - X\beta)\right]\right). \end{aligned}$$

The density on the last page is easily recognized as the kernel of an **inverse gamma** density:

$$IG\left(\frac{n}{2} + a, \left[b^{-1} + \frac{1}{2}(y - X\beta)'(y - X\beta)\right]^{-1}\right)$$

density.

- Drawing from the Inverse Gamma is easy - if a routine for the generation of Gamma variates is available, you can simply invert them (but be careful about parameterization!!)

Thus, to implement the Gibbs sampler in the linear regression model, we can proceed as follows.

- 1 Given a current value of  $\sigma^2$  :
- 2 Calculate  $D_\beta = D_\beta(\sigma^2)$  and  $d_\beta = d_\beta(\sigma^2)$ .
- 3 Draw  $\tilde{\beta}$  from a  $N[D_\beta(\sigma^2)d_\beta(\sigma^2), D_\beta(\sigma^2)]$  distribution.
- 4 Draw  $\tilde{\sigma}^2$  from an

$$IG\left(\frac{n}{2} + a, \left[b^{-1} + \frac{1}{2}(y - X\tilde{\beta})'(y - X\tilde{\beta})\right]^{-1}\right)$$

distribution.

- 5 Repeat this process many times, updating the posterior conditionals at each iteration to condition on the most recent simulations produced in the chain.
- 6 Discard an early set of parameter simulations as the burn-in period.
- 7 Use the subsequent draws to compute posterior features of interest.

## Application to Wage Data

- We apply this Gibbs sampler using a small sample of 1985 outcomes from the NLSY.
- We consider the model

$$\log Wage_i = \beta_0 + \beta_1 Education_i + \epsilon_i, \quad \epsilon_i | Ed \sim N(0, \sigma^2 I_n).$$

- The following estimation results are obtained under the prior

$$\beta \sim N(0, 4I_2), \quad \sigma^2 \sim IG[3, (1/[2 * .2])].$$

- We obtain 5,000 simulations, and discard the first 100 as the burn-in period.

## Posterior Calculations Using the Gibbs Sampler

	$\beta_0$	$\beta_1$	$\sigma^2$
$E(\cdot y)$	1.18	.091	.267
$Std(\cdot y)$	.087	.0063	.0011

- Posterior means are virtually indistinguishable from OLS estimates.
- In addition, we calculate, say,

$$\Pr(\beta_1 < .10|y) = .911$$

to illustrate how the simulations can be used to calculate a variety of quantities of interest (and to contrast interpretations with the frequentist case).

Posterior Prediction

The **method of composition** can also prove to be a very valuable tool for problems of (posterior) prediction.

To this end, consider an out-of-sample value  $y_f$  which is presumed to be generated by our regression model:

$$y_f = X_f\beta + u_f, \quad u_f|X_f \sim N(0, \sigma^2).$$

- 1 Note that  $y_f|\beta, \sigma^2$  does not depend on  $y$ . (But does through  $\beta$  and  $\sigma^2$ .)
- 2 The goal is to simulate draws from the posterior predictive:

$$p(y_f|y),$$

which does not depend on any of the model's parameters.

Posterior Prediction

To generate draws from this posterior predictive, we first consider the joint posterior distribution:

$$p(y_f, \beta, \sigma^2|y).$$

If we can draw from this distribution, we can use only the  $y_f$  draws (and ignore those associated with  $\beta$  and  $\sigma^2$ ) as draws from the marginal  $p(y_f|y)$ .

How can we do this?

Note

$$p(y_f, \beta, \sigma^2|y) = p(y_f|\beta, \sigma^2, y)p(\beta, \sigma^2|y)$$

This suggests that draws from the marginal can be obtained by

- 1 Drawing a  $(\tilde{\sigma}^2, \tilde{\beta})$  from  $p(\beta, \sigma^2|y)$ . [We have these from the Gibbs Sampler].
- 2 Drawing  $y_f$  from a  $N(X_f\tilde{\beta}, \tilde{\sigma}^2)$  distribution.
- 3 Note, of course, this requires that  $X_f$  is known.
- 4 Doing this many times will produce a set of draws from the posterior predictive  $y_f|y$ .

- Let's apply this method to generate draws from the posterior predictive using our log wage example:

$$\log(\text{wage})_i = \beta_0 + \beta_1 \text{Education}_i + u_i.$$

- The method just described could be applied directly to sample from the predictive distribution of (log) hourly wages.
- However, the *wage density* itself is actually more interpretable.
- To sample from the posterior predictive of wages (in levels), we can consider drawing from an augmented density of the form:

$$p(w_f, y_f, \beta, \sigma^2 | y)$$

where

$$w_f = \exp(y_f).$$

$$p(w_f, y_f, \beta, \sigma^2 | y)$$

We can write this joint distribution as follows:

$$\begin{aligned} p(w_f, y_f, \beta, \sigma^2 | y) &= p(w_f, y_f | \beta, \sigma^2, y) p(\beta, \sigma^2 | y) \\ &= p(w_f | y_f, \beta, \sigma^2, y) p(y_f | \beta, \sigma^2, y) p(\beta, \sigma^2 | y) \\ &= p(w_f | y_f) p(y_f | \beta, \sigma^2) p(\beta, \sigma^2 | y) \end{aligned}$$

where the last line follows since the distribution of  $w_f$  only depends on  $y_f$  and, in fact,

$$p(w_f | y_f) = I[w_f = \exp(y_f)].$$

Thus, within the context of our example, we can generate draws from the posterior predictive distribution of *hourly wages*  $w_f$  as follows:

- 1 Generate  $(\tilde{\sigma}^2, \tilde{\beta})$  from  $p(\beta, \sigma^2 | y)$ .

- 2 Generate

$$y_f \sim N(X_f \tilde{\beta}, \tilde{\sigma}^2).$$

- 3 Calculate

$$w_f = \exp(y_f).$$

- We apply this technique to our data set and generate 10,000 draws from the posterior predictive distribution of hourly wages for two cases:  $\text{Ed} = 12$  and  $\text{Ed} = 16$ .
- The 10,000 draws are then smoothed nonparametrically via a kernel density estimator.
- Graphs of these densities are provided on the following page.

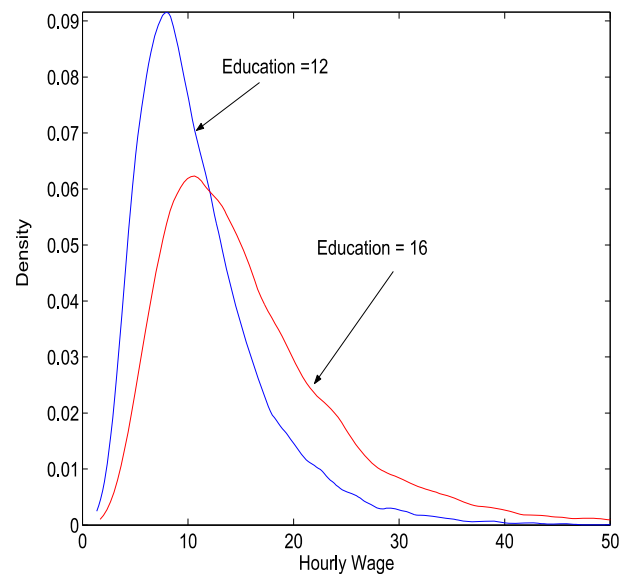


Figure : Posterior Predictive Hourly Wage Densities

- The (posterior predictive) mean hourly wage for high school graduates is (approximately) \$11.07.
- The mean hourly wage for those with a BA is (approximately) \$15.88
- The posterior probability that a high school graduate will receive an hourly wage greater than \$15 is

$$\Pr(w_f > 15 | Ed_f = 12, y) \approx .19$$

- The posterior probability that an individual with a BA will receive an hourly wage greater than \$15 is

$$\Pr(w_f > 15 | Ed_f = 16, y) \approx .44$$

- If you are curious, doing the same exercise for someone with a Ph.D., i.e.,  $Ed = 20$ , gives  $\Pr(w_f > 15 | Ed_f = 20, y) \approx .72$

Don't worry - this is old data!!!

## Gibbs Sampling in a Probit



### Probit Model

Consider the following latent variable representation of the probit model:

$$z_i = x_i\beta + \epsilon_i, \quad \epsilon_i | X \stackrel{iid}{\sim} N(0, 1),$$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}.$$

The value of the binary variable  $y_i$  is observed, as are the values of the explanatory variables  $x_i$ . The latent data  $z_i$ , however, are unobserved.

- Tanner and Wong (1987) and Albert and Chib (1993) describe the idea of **data augmentation**.
  -  Tanner, M. and W. Wong (1987). 'The Calculation of Posterior Distributions by Data Augmentation'. *Journal of the American Statistical Association*, 82: 528-49.
  -  Albert, J. and S. Chib (1993). 'Bayesian Analysis of Binary and Polychotomous Response Data'. *Journal of the American Statistical Association*, 88: 669-79.
- For many microeconomic models the idea is that, **conditioned on suitably defined latent data**, the models are effectively linear, so all of the usual Lindley-Smith (1972) mechanics apply.
- That is, instead of working with  $p(\theta|y)$ , we might try to implement a Gibbs algorithm on  $p(\theta, z|y)$  since  $p(\theta|z, y)$  might be much easier to deal with than  $p(\theta|y)$ .

To derive the *augmented* joint posterior for the probit, note that

$$p(\beta, z|y) = \frac{p(y, z|\beta)p(\beta)}{p(y)},$$

implying

$$p(\beta, z|y) \propto p(y, z|\beta)p(\beta).$$

- The term  $p(\beta)$  is simply our prior, while  $p(y, z|\beta)$  represents the **complete** or **augmented** data density.

To characterize this density in more detail, note

$$p(y, z|\beta) = p(y|z, \beta)p(z|\beta).$$

Immediately, from our latent variable representation, we know

$$p(z|\beta) = \prod_{i=1}^n \phi(z_i; x_i\beta, 1).$$

As for the conditional for  $y$  given  $z$  and  $\beta$ , note that when  $z_i > 0$  then  $y_i$  must equal one, while when  $z_i \leq 0$ , the  $y_i$  must equal zero.

In other words, the sign of  $z_i$  perfectly predicts the value of  $y$ . Hence, we can write

$$p(y|z, \beta) = \prod_{i=1}^n [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)].$$

Putting the pieces together, we obtain the augmented data density  $p(y, z|\beta)$ . We combine this with our prior to obtain

$$p(\beta, z|y) \propto p(\beta) \prod_{i=1}^n [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)] \phi(z_i, x_i\beta, 1).$$

It might be useful to see what we get when we integrate out the latent data:

$$\begin{aligned}
 p(\beta|y) &= \int_z p(\beta, z|y) dz \\
 &\propto p(\beta) \int_z \prod_{i=1}^n [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)] \phi(z_i, x_i\beta, 1) dz_1 \cdots dz_n \\
 &= p(\beta) \prod_{i=1}^n \left[ \int_{-\infty}^{\infty} [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)] \phi(z_i, x_i\beta, 1) dz_i \right] \\
 &= p(\beta) \prod_{i=1}^n \left[ \int_{-\infty}^0 I(y_i = 0) \phi(z_i; x_i\beta, 1) dz_i + \int_0^{\infty} I(y_i = 1) \phi(z_i; x_i\beta, 1) dz_i \right] \\
 &= p(\beta) \prod_{i=1}^n [I(y_i = 0)[1 - \Phi(x_i\beta)] + I(y_i = 1)\Phi(x_i\beta)] \\
 &= p(\beta) \prod_{i=1}^n \Phi(x_i\beta)^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i}.
 \end{aligned}$$

- Note that this is exactly the prior times the likelihood that one would produce without the introduction of any latent variables.
- What is important to note is that the posterior of  $\beta$  is unchanged by the addition of the latent variables.
- So, augmenting the posterior with  $z$  will not change any inference regarding  $\beta$  - it is the same as if we would have worked with the (nonlinear) likelihood directly - though it does make the problem computationally easier, as we will see below.

Suppose a prior of the following form is employed:

$$\beta \sim N(\mu_\beta, V_\beta).$$

The complete conditional for  $\beta$  *given*  $z$  and the data  $y$  follows directly from standard results from the linear regression model

$$\beta|z, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = (X'X + V_\beta^{-1})^{-1}, \quad d_\beta = X'z + V_\beta^{-1}\mu_\beta.$$

As for the complete conditional for  $z$ , first note that the independence across observations implies that each  $z_i$  can be drawn independently.

We also note that

$$\begin{aligned}
 z_i|\beta, y &\propto I(z_i > 0)\phi(z_i; x_i\beta, 1) & \text{if } y_i = 1 \\
 z_i|\beta, y &\propto I(z_i \leq 0)\phi(z_i; x_i\beta, 1) & \text{if } y_i = 0.
 \end{aligned}$$

Thus,

$$z_i | \beta, y \stackrel{\text{iid}}{\sim} \begin{cases} \text{TN}_{(-\infty, 0]}(x_i \beta, 1) & \text{if } y_i = 0 \\ \text{TN}_{(0, \infty)}(x_i \beta, 1) & \text{if } y_i = 1 \end{cases}.$$

- The notation  $\text{TN}_{[a, b]}(\mu, \sigma^2)$  denotes a Normal distribution with mean  $\mu$  and variance  $\sigma^2$  truncated to the interval  $[a, b]$ .
- To generate draws from the Truncated Normal one can use the method of inversion - there are all kinds of m-files or R programs out there that do this!
- So, to fit a probit, you just need to be able to sample from a multivariate normal and univariate truncated normal.

## Ordered Probit Model

Consider a latent variable representation of the **ordered probit model**:

$$z_i = x_i \beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1),$$

and

$$y_i = \begin{cases} 1 & \text{if } \alpha_0 < z_i \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < z_i \leq \alpha_2 \\ \vdots & \vdots \\ M & \text{if } \alpha_{M-1} < z_i \leq \alpha_M \end{cases}.$$

The latent variable  $z$  is not observed.

For identification purposes, we set  $\alpha_0 = -\infty$ ,  $\alpha_1 = 0$  and  $\alpha_M = \infty$ . The  $\alpha_i$  are often called **cutpoints**.

## Ordered Probit Model

- It is well-known in the literature that, particularly in reasonably large data sets, standard Gibbs in the ordered probit suffers from slow mixing.
- Part of the reason for this slow mixing is the result of high correlation between the simulated cutpoints  $\alpha$  and latent data  $z$ .
- Cowles (1996), for example, recommends employing a blocking step (where the  $\alpha$  and  $z$  are drawn together in a single block.) This blocking procedure uses an M-H step to sample the cutpoints from a series of truncated normal proposal densities.

## Reparameterization in the Ordered Probit

Here, we describe an alternate reparameterization strategy as suggested by Nandram and Chen (1996).

To fix ideas, let us consider the case where  $M = 3$  and so there is only one unknown cutpoint  $\alpha_2$ .



Cowles, M.K. (1996). "Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models." *Statistics and Computing* 6, 101-111.



Nandram, B. and M-H Chen (1996). "Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence." *Journal of Statistical Computation and Simulation*.

Let us take our original ordered probit:

$$z_i = x_i \beta + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, 1),$$

$$y_i = \begin{cases} 1 & \text{if } -\infty < z_i \leq 0 \\ 2 & \text{if } 0 < z_i \leq \alpha_2 \\ 3 & \text{if } \alpha_2 < z_i \leq \infty \end{cases}$$

and multiply the latent equation by  $\delta = 1/\alpha_2$ .

We then obtain the following, observationally equivalent model:

$$z_i^* = x_i \beta^* + \epsilon_i^*, \quad \epsilon_i^* \stackrel{iid}{\sim} N(0, \delta^2),$$

$$y_i = \begin{cases} 1 & \text{if } -\infty < z_i^* \leq 0 \\ 2 & \text{if } 0 < z_i^* \leq 1 \\ 3 & \text{if } 1 < z_i^* \leq \infty \end{cases}$$

In this reparameterized model, **there are no unknown cutpoints** and we work with the parameterization

$$\delta = 1/\alpha_2, \quad \beta^* = \delta \beta, \quad z^* = \delta z.$$

- So, when  $M = 3$ , we can simply use a standard Gibbs algorithm, as if the model were a (latent) linear regression model.
- When  $M > 3$ , we must sample the remaining cutpoints. One possibility is to sample differences in cutpoint values via a Dirichlet proposal density [e.g., Nandram and Chen (1996)].

## Multinomial Probit Model

Suppose that an agent has a choice among  $J$  alternatives, with no natural ordering among the alternatives.

Let  $y_i$  denote the observed choice of the agent, with  $y_i \in \{0, 1, \dots, J-1\}$ .

Let  $U_{ij}$  represent the latent utility received by agent  $i$  from making choice  $j$ . We assume

$$U_{ij} = x_{ij} \beta_j + \epsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 0, 1, \dots, J-1.$$

Note that we can difference with respect to the base category (0) and stack the observations over  $j$  for each  $i$  as follows: (where  $z$  and  $\beta$  can be modified as needed):

$$\begin{bmatrix} U_{i1}^* \\ U_{i2}^* \\ \vdots \\ U_{iJ-1}^* \end{bmatrix} = \begin{bmatrix} z_{i1} & 0 & \cdots & 0 \\ 0 & z_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_{iJ-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{J-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1}^* \\ \epsilon_{i2}^* \\ \vdots \\ \epsilon_{iJ-1}^* \end{bmatrix},$$

or equivalently,

$$U_i^* = Z_i \beta + \epsilon_i^*, \quad \epsilon_i^* \stackrel{iid}{\sim} N(0, \Sigma),$$

(where a scale normalization is also required).

We work with the augmented joint posterior of the form

$$p(U^*, \beta, \Sigma^{-1} | y),$$

where

$$U^* = \begin{bmatrix} U_1^* \\ U_2^* \\ \vdots \\ U_n^* \end{bmatrix}.$$

This posterior distribution can be expressed as follows





$$p(U^*, \beta, \Sigma^{-1} | y) \propto p(\beta) p(\Sigma^{-1}) \left[ \prod_{i=1}^n \phi(U_i^*; z_i \beta, \Sigma) \left( I(y_i = 0) I(\max\{U_{ij}^*\} \leq 0) \right. \right. \\ \left. \left. + \sum_{k=1}^{J-1} I(y_i = k) I[U_{ik}^* > \max\{0, U_{i-k}^*\}] \right) \right],$$

with  $U_{i-k}^*$  denoting the collection of utility differences for agent  $i$  other than the  $k^{th}$  difference.

- In the first Bayesian work on this model, McCulloch and Rossi (1994) fit the model using an “unrestricted”  $\Sigma$  and simply report posteriors for identifiable parameters by normalizing with respect to a diagonal element:

$$\tilde{\beta} = \beta / \sqrt{\sigma_{11}}, \quad \tilde{\Sigma} = \Sigma / \sigma_{11}.$$

- McCulloch et al (2000) reparameterize  $\Sigma$  and describe a (Gibbs) posterior simulator over the identifiable parameter space.
- However, the prior over the unidentified model induces a prior on the identified parameters that is not in standard form, and it is not in general easy to elicit sensible hyperparameters in this context (see Imai and van Dyk (2005 *JoE*).

-  McCulloch, R. and P.E. Rossi (1994). “An exact likelihood analysis of the multinomial probit model” *Journal of Econometrics* 64, 207-240.
-  McCulloch, R., N. Polson and P.E. Rossi (2000). “A Bayesian analysis of the multinomial probit model with fully identified parameters” *Journal of Econometrics* 99, 173-193.
-  Nobile, A. (2000). “Comment: Bayesian multinomial probit models with a normalization constraint ” *Journal of Econometrics* 99, 335-345.
-  Imai, K. and van Dyk, D. A. (2005). 11A Bayesian Analysis of the Multinomial Probit Model Using Marginal Augmentation.” *Journal of Econometrics*, 311-334.

## Putting Things Together: A LRM with an Endogenous Dummy

### A Standard Binary Treatment Model

Consider the following model with a continuous outcome  $y$  and a binary treatment variable  $D$ :

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 D_i + \epsilon_i \\ D_i^* &= z_i \theta + u_i, \end{aligned}$$

where

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \Big| z \stackrel{iid}{\sim} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{u\epsilon} \\ \sigma_{u\epsilon} & 1 \end{pmatrix} \right] \equiv N(0, \Sigma).$$

- The second equation of the system describes a latent variable generating  $D$ , i.e.,  $D = I(D^* > 0)$ , like the probit model previously discussed.
- We seek to describe a Gibbs sampling algorithm for fitting this two equation system.

### A Standard Binary Treatment Model

First, note that we can stack the model into the form

$$\tilde{y}_i = X_i \beta + \tilde{u}_i,$$

where

$$\tilde{y}_i = \begin{bmatrix} y_i \\ D_i^* \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 & D_i & 0 \\ 0 & 0 & z_i \end{bmatrix}, \quad \beta = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \theta \end{bmatrix}, \quad \tilde{u}_i = \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix}.$$

- In this form, posterior simulation seems very similar to a SUR model [Given the unit Jacobian of the transformation].
- However, the fact that the (2,2) element of the covariance matrix must be restricted to unity introduces some complications. In addition, we must sample  $D^*$ .

### A Standard Binary Treatment Model

We suggest a reparameterization to help with the former issue. First, write:

$$\epsilon_i = \sigma_{u\epsilon} u_i + v_i,$$

where  $v_i \sim N(0, \sigma_v^2)$ ,  $v$  and  $u$  are independent, and  $\sigma_v^2 \equiv \sigma_\epsilon^2 - \sigma_{u\epsilon}^2$ .

So, we can work with an equivalent version of the model:

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 D_i + \sigma_{u\epsilon} u_i + v_i \\ D_i^* &= z_i \theta + u_i, \end{aligned}$$

where  $u$  and  $v$  are **independently distributed**. In this parameterization,  $\Sigma$  takes the form:

$$\Sigma = \begin{bmatrix} \sigma_v^2 + \sigma_{u\epsilon}^2 & \sigma_{u\epsilon} \\ \sigma_{u\epsilon} & 1 \end{bmatrix}.$$

## A Standard Binary Treatment Model

We complete the model by choosing priors of the following forms:

$$\begin{aligned}\beta &\sim N(\mu_\beta, V_\beta) \\ \sigma_{u\epsilon} &\sim N(\mu_0, V_0) \\ \sigma_v^2 &\sim IG(a, b),\end{aligned}$$

Finally, note that  $\Sigma$  is positive definite for  $\sigma_v^2 > 0$ , which is enforced through our prior.

## A Standard Binary Treatment Model

We work with an augmented posterior distribution of the form

$$p(\beta, D^*, \sigma_v^2, \sigma_{u\epsilon} | y, D).$$

As for the complete conditional for  $\beta$ , its derivation follows identically to the SUR model:

$$\beta | D^*, \Sigma, y, D \sim N(D_\beta d_\beta, D_\beta)$$

where

$$D_\beta = (\sum_i X_i' \Sigma^{-1} X_i + V_\beta^{-1})^{-1}, \quad d_\beta = \sum_i X_i' \Sigma^{-1} \tilde{y}_i + V_\beta^{-1} \mu_\beta.$$

(Note that  $\Sigma$  is known given  $\sigma_v^2$  and  $\sigma_{u\epsilon}$ ).

As for the posterior conditional for each  $D_i^*$ , we must first break our likelihood contributions into a conditional for  $D_i^* | y_i$  and a marginal for  $y_i$ . Thus,

$$D_i^* | \beta, \Sigma, y, D \stackrel{ind}{\sim} \begin{cases} TN_{(0, \infty)}[z_i \theta + [\sigma_{u\epsilon} / (\sigma_v^2 + \sigma_{u\epsilon}^2)](y_i - \alpha_0 - \alpha_1), (1 - \rho_{u\epsilon}^2)] & : D_i = 1 \\ TN_{(-\infty, 0]}[z_i \theta + [\sigma_{u\epsilon} / (\sigma_v^2 + \sigma_{u\epsilon}^2)](y_i - \alpha_0), (1 - \rho_{u\epsilon}^2)] & : D_i = 0. \end{cases}$$

As for the parameters of the covariance matrix, let us go back to our earlier version of the model:

$$\begin{aligned}y_i &= \alpha_0 + \alpha_1 D_i + \sigma_{u\epsilon} u_i + v_i \\ D_i^* &= z_i \theta + u_i,\end{aligned}$$

Note that, conditioned on  $\theta$  and  $D^*$ , the errors  $u$  are “known” and thus we can treat  $u$  as a typical regressor in the first equation when sampling from the posterior conditional for  $\sigma_{u\epsilon}$ :

$$\sigma_{u\epsilon} | \beta, D^*, \sigma_v^2, y, D \sim N(Dd, D)$$

where

$$D = (u' u / \sigma_v^2 + V_0^{-1})^{-1}, \quad d = u'(y - \alpha_0 - \alpha_1 D) / \sigma_v^2 + V_0^{-1} \mu_0.$$

Finally,

$$\sigma_v^2 | \sigma_{ue}, \beta, D^*, y, D \sim IG \left( \frac{n}{2} + a, [b^{-1} + .5 \sum_i (y_i - \alpha_0 - \alpha_1 D_i - \sigma_{ue} u_i)^2]^{-1} \right).$$

- The Gibbs sampler proceeds by cycling through all of these conditionals, and it is easy to simulate draws from each of these.
- At each iteration, we can calculate the structural parameter  $\sigma_\epsilon^2 = \sigma_v^2 + \sigma_{ue}^2$ .

## Flexible Extensions

- All of the analysis to this point assumed Gaussian errors.
- This may not be appropriate and people may not be willing to give you a job / tenure if you hang your hat on normality too often.
- So, is there a way to maintain the computational conveniences of Gaussian sampling models (under Gaussian priors) yet be flexible and able to adapt to handle non-Gaussian situations?

## Flexible Extensions

- A simple extension is through **scale mixtures of normals**.
- For example, if we specify

$$\epsilon | \lambda, \sigma^2 \sim N(0, \lambda \sigma^2), \quad \lambda \sim IG \left( \frac{\nu}{2}, \frac{2}{\nu} \right)$$






then

$$\epsilon | \sigma^2 \sim t(0, \sigma, \nu)$$

(again, be careful about parameterization)

- In terms of Gibbs implementation, then, we can simply add an additional step to sample the  $\lambda$  mixing variables (which turn out to be conditionally inverse Gamma) to fit a model that has Student-t errors.
- Other mixing distributions give rise to other marginals for  $\epsilon$  (e.g., logit, double exponential).

## A Few References on Scale Mixtures

-  Andrews, D.F. and C.L. Mallows (1974). 'Scale Mixtures of Normal Distributions'. *Journal of the Royal Statistical Society, Series B*, 36: 99-102.
-  Carlin, B.P. and N. G. Polson (1991). 'Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler'. *The Canadian Journal of Statistics*, 19: 399-405.
-  Geweke, J. (1993). 'Bayesian Treatment of the Independent Student-t Linear Model'. *Journal of Applied Econometrics*, 8: S19-S40.
-  Albert, J. and S. Chib (1993). 'Bayesian Analysis of Binary and Polychotomous Response Data'. *Journal of the American Statistical Association*, 88: 669-79.
-  Chib, S. and B. Hamilton (2000). 'Bayesian Analysis of Cross Section and Clustered Data Treatment Models'. *Journal of Econometrics*, 97: 25-50.

- An alternate approach is to employ **finite mixtures**.
- For example, let  $r$  denote a  $J \times 1$  component label vector. Given  $r$ , we could consider a model of the form:

$$y_i = \left( \sum_{j=1}^J \alpha_{0j} I(r_i = j) \right) + \alpha_1 D_i + \epsilon_i$$

$$D_i^* = \left( \sum_{j=1}^J \theta_{0j} I(r_i = j) \right) + z_i \tilde{\theta} + u_i$$

where

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \Big| r, z \stackrel{\text{ind}}{\sim} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\epsilon}^{2,(r)} & \sigma_{\epsilon u}^{(r)} \\ \sigma_{\epsilon u}^{(r)} & 1 \end{pmatrix} \right] \equiv N(0, \Sigma_r).$$

- So, the error covariance matrix and intercept parameters will vary across the mixture components.
- The model is completed by specifying


$$\Pr(r_i = j | p) = p_j, \quad \sum_{j=1}^J p_j = 1, \quad i = 1, 2, \dots, n$$

and adding a Dirichlet prior over the probability vector  $p$ .


- This allows a great deal of flexibility. Sometimes the mixtures can be used / argued to capture various forms of heterogeneity.
- In terms of posterior simulation, **conditioned on  $r$** , the model is just like the one we considered previously, and thus standard Gibbs can be applied.
- Extra steps are added for the sampling of  $r$  (Multinomial) and  $p$  (Dirichlet).

## Just a Few of Many Papers on Mixtures / Flexible Models

- An even greater degree of flexibility can be obtained by using a Dirichlet process prior, as in Conley et al (2008).


 Li, M., D.J. Poirier and J.L. Tobias (2004). 'Do Dropouts Suffer from Dropping Out? Estimation and Prediction of Outcome Gains in Generalized Selection Models'. *Journal of Applied Econometrics*, 19: 203-25.

 Smith, M.D. (2005). "Using Copulas to Model Switching Regimes with an Application to Child Labor" *Economic Record*, S47-S57.

 Geweke, J. and M. Keane (2007). 'Smoothly Mixing Regressions'. *Journal of Econometrics*, 138: 252-91.

 Conley, T., C. Hansen, R. McCulluch and P. Rossi (2008). 'A Semi-Parametric Bayesian Approach to the Instrumental Variables Problem'. *Journal of Econometrics*, 144: 276-305.

 Villani, M. R. Kohn and P. Giordani (2009). "Regression density Estimation Using Smooth Adaptive Gaussian Mixtures" *Journal of Econometrics*, 155-173.

 Griffin, J., F. Quintana and M.F.J. Steel (2011). 'Flexible and Nonparametric Modelling', in J. Geweke, G. Koop and H. van Dijk, (eds.), *Handbook of Bayesian Econometrics*. Oxford: Oxford University Press, 125-182.

 Villani, M. R. Kohn and D. J. Nott (2012). "Generalized Smooth Finite Mixtures" *Journal of Econometrics*, forthcoming.

## Flexible Regression Functions: Kline and Tobias (2008)

$$y_i = f(s_i) + \mathbf{x}_i \beta + \epsilon_i$$

$$s_i = \mathbf{z}_i \theta + \delta h_i^* + u_i,$$

where

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \Big| \mathbf{x}, \mathbf{z} \stackrel{iid}{\sim} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\epsilon}^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{pmatrix} \right] \quad i = 1, 2, \dots, n.$$

- The endogenous variable  $s$  is treated **nonparametrically** in the outcome equation.
- $h_i^*$  are latent variables, (e.g. *iid* from a half normal), to allow for additional skew in the conditional BMI distribution

## Bayesian Implementation

Stack the observations over  $i$  to obtain:

$$\begin{aligned} \mathbf{y} &= \mathbf{D}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \mathbf{s} &= \mathbf{Z}\boldsymbol{\theta} + \delta\mathbf{h}^* + \mathbf{u} \end{aligned}$$

where

$$[\boldsymbol{\epsilon}' \mathbf{u}']' | \mathbf{x}, \mathbf{z}, \mathbf{h}^* \sim N(\mathbf{0}_{2n}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n), \quad \boldsymbol{\gamma} \equiv \begin{bmatrix} f(s_1) \\ f(s_2) \\ \vdots \\ f(s_{k_\alpha}) \end{bmatrix},$$

$D$  is a  $n \times k_\gamma$  matrix,  $k_\gamma \leq n$ , with  $i$ th row  $\mathbf{d}_i$ , which is constructed to select off the appropriate element of  $\boldsymbol{\gamma}$ , and  $s_j < s_{j+1}$ .

## Bayesian Implementation

To smooth the curve, we reparameterize the model in terms of  $\boldsymbol{\psi} \equiv \mathbf{H}\boldsymbol{\gamma}$ , where

$$\mathbf{H} \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ \Delta_2^{-1} & -\Delta_2^{-1} - \Delta_3^{-1} & \Delta_3^{-1} & 0 & \dots & 0 & 0 & 0 \\ 0 & \Delta_3^{-1} & -\Delta_3^{-1} - \Delta_4^{-1} & \Delta_4^{-1} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \Delta_{k_\alpha-1}^{-1} & -\Delta_{k_\alpha-1}^{-1} - \Delta_{k_\alpha}^{-1} & \Delta_{k_\alpha}^{-1} \end{bmatrix},$$

and  $\Delta_j \equiv s_j - s_{j-1}$ .

The elements of  $\boldsymbol{\psi} = \mathbf{H}\boldsymbol{\gamma}$  thus consist of a pair of “initial conditions”  $\gamma_1$  and  $\gamma_2$  and differences of the form

$$\begin{aligned} \psi_j &= \frac{\gamma_j - \gamma_{j-1}}{s_j - s_{j-1}} - \frac{\gamma_{j-1} - \gamma_{j-2}}{s_{j-1} - s_{j-2}}, \quad j = 3, 4, \dots, k_\gamma \\ &\approx f'(s_{j-1})' - f'(s_{j-2}). \end{aligned}$$

## Bayesian Implementation

We choose a prior for  $\boldsymbol{\psi}$  of the form:

$$[\psi_1 \ \psi_2]' \sim N[\mathbf{0}_2, 10\mathbf{I}_2],$$

with  $\mathbf{I}_2$  denoting the  $2 \times 2$  identity matrix.

For the remaining elements of  $\boldsymbol{\psi}$ , we specify a prior of the form

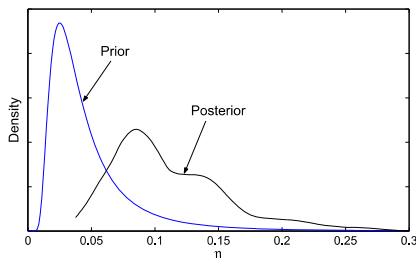
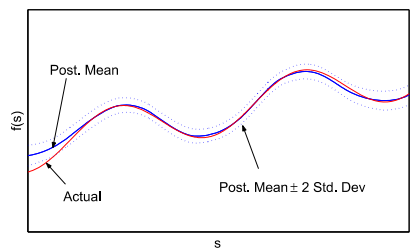
$$[\psi_3 \ \psi_4 \ \dots \ \psi_{k_\alpha}]' | \eta \sim N[\mathbf{0}_{k_\alpha-2}, \eta \mathbf{I}_{k_\alpha-2}],$$

- As  $\eta \rightarrow 0$ , force **linearity**.
- For “intermediate”  $\eta$ , force **local similarity**.
- Allow  $\eta$  to be updated by the data.
- Can also use Gaussian Process priors to the same end.

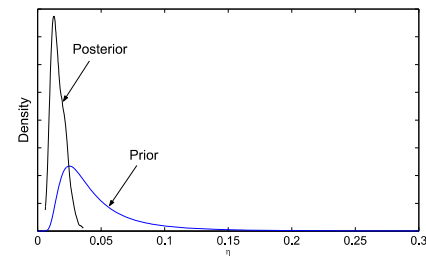
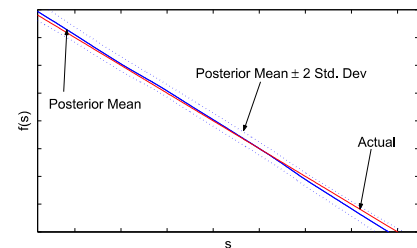
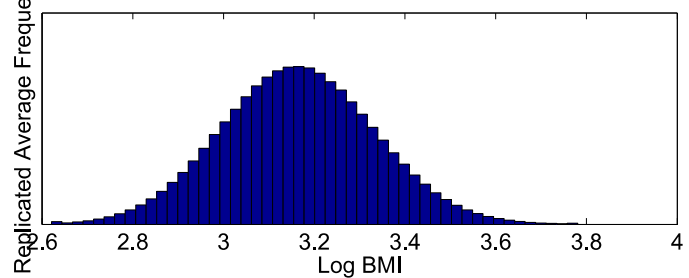
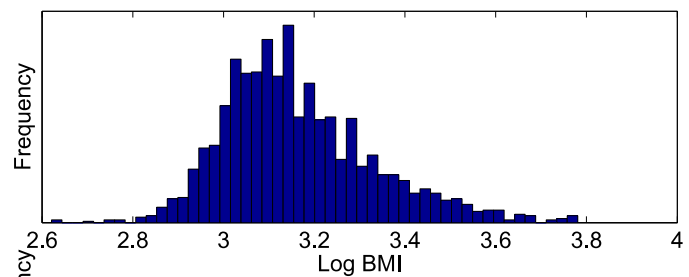
## Generated Data Experiments

- Generate some data from the model described above to investigate the performance of the algorithm.
- Do this for two cases. One where  $f$  is nonlinear (a sin function with a linear trend), and the second for a linear model.
- Keep priors constant in both experiments.
- Focus on posterior results for  $f(s)$  and the smoothing parameter  $\eta$  to fix ideas.

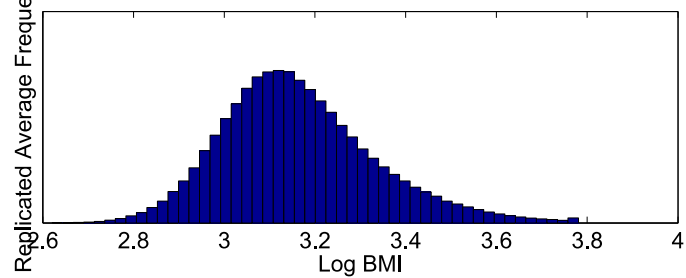
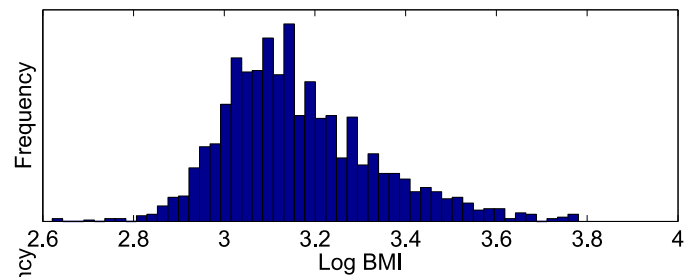
## Generated Data Results: Nonlinear Model



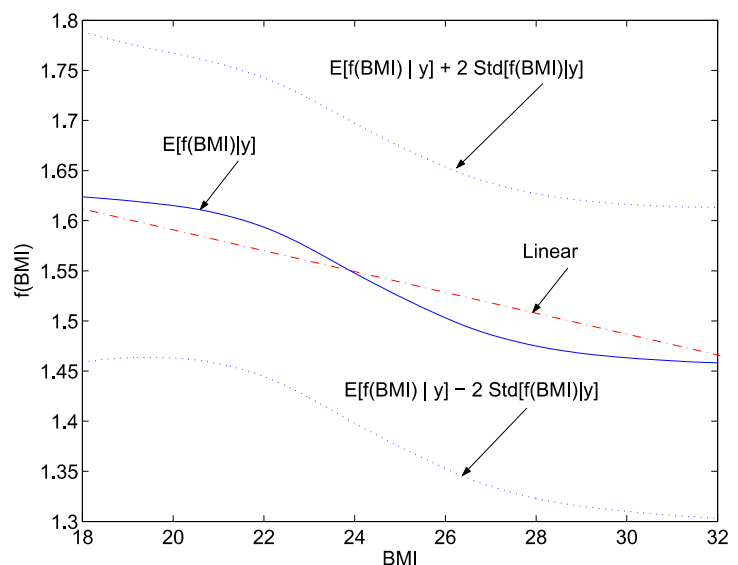
## Generated Data Results: Linear Model

Diagnostics with Actual Data, Using  $s = \log(BMI)$ 

## Diagnostics with Actual Data, Assuming Skew-Normality



## Regression function $f(s)$ : Females Sample



## IV Imperfection

$$y_i = \alpha_0 + z_i\alpha_1 + \mathbf{x}_i\alpha_2 + \gamma s_i + u_i,$$

$$s_i = \beta_0 + z_i\beta_1 + \mathbf{x}_i\beta_2 + v_i,$$

where

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \middle| \mathbf{X}, \mathbf{z} \stackrel{iid}{\sim} N \left[ \mathbf{0}, \begin{pmatrix} \sigma_u^2 & \rho_{uv}\sigma_u\sigma_v \\ \rho_{uv}\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} \right].$$

- $y$  is the continuous outcome,  $s$  is a continuous endogenous variable.
- $\gamma$  is typically the key parameter of interest.
- The model is only **partially identified**.

## The Model

- $z$  is differentiated from  $x$  in that  $z$  is potentially excludable.
- The dominant approach in practice is to employ  $p(\alpha_1) = \mathbf{1}(\alpha_1 = 1)$  and then spend considerable time and effort to try and convince us that this prior is “correct.”
- One can employ different priors for  $\alpha_1$  than we will for elements of  $\alpha_2$  given our knowledge of the problem at hand, yet we do not need to impose perfect excludability.
- Conley, Hansen and Rossi (2012) consider

$$\alpha_1 \sim N(\mu, \delta^2), \quad \alpha_2 \sim N(\mu, \gamma^2 \delta^2).$$



Conley, T.G., C.B. Hansen and P.E. Rossi (2012). “Plausibly Exogenous” *The Review of Economics and Statistics*, 260-272.

- At a minimum, one can conduct an analysis of this sort as a robustness exercise - to illustrate how posterior results change as prior information changes.
- This is consistent with what is being done in the bounding literature.
- Bayesians often support the production of a menu of posterior results; in this particular case, dogmatic priors correspond to just a single item on this menu.

## IV Imperfection


- Kraay (2012) notes posterior sensitivity to even mild uncertainty surrounding excludability.
- Chan and Tobias (2012) discuss posterior computation in the partially identified setting.
  - They consider priors that order “direct” and “indirect” effects, e.g:

$$\alpha_1 | \gamma, \beta_1 \sim TN_{(-|\beta_1 \gamma|, |\beta_1 \gamma|)}(0, V_{\alpha_1}).$$

- Non-ID coefficients mix very poorly, and NSEs will be relatively large by necessity.
- Offer a computational approach that improves upon iid simulation from the joint posterior.

**Table :** Number of draws (in thousands) required to have the NSE less than one percent of the posterior mean:  $\alpha_{1i} \sim N(0, .05^2)$ .

Wage Equation					
Parameter	$E(\cdot   Data)$	Number of Iterations		Computation Time	
		Gibbs	Semi-Analytic	Gibbs	Semi-Analytic
BMI	.008	3,377,208	146	390 days	24 minutes
Constant	1.756	5,685	1.50	15.8 hours	.25 minutes
MomBMI	-.0092	363,582	8	42.1 days	1.2 minutes
DadBMI	-.0073	317,180	24	36.7 days	4.0 minutes
FamilyIncome	.0007	218	63	0.6 hours	10 minutes
HighSchool	.062	5,659	14	15.7 hours	2.4 minutes
Alevel	.266	899	1.92	2.5 hours	.32 minutes
Degree	.355	903	.58	2.5 hours	.09 minutes
Union	.031	6,852	38	19.0 hours	6.3 minutes
Married	-.018	55,708	106	6.4 days	17 minutes
Other Parameters					
Parameter	$E(\cdot   Data)$	Gibbs	Semi-Analytic	Gibbs	Semi-Analytic
$\rho_{uv}$	-.097	965,848	5	112 days	.90 minutes
$\sigma_u^2$	.225	12,293	.03	1.4 days	.005 minutes

-  Poirier, D.J. (1998). “Revising Beliefs in Non-Identified Models” *Econometric Theory*, 483-509.
-  Nevo, A. and A.M. Rosen (2012). “Identification with Imperfect Instruments” *Review of Economics and Statistics*, 659-671.
-  Kraay, A. (2012). “Instrumental Variables Regressions with Uncertain Exclusion Restrictions: A Bayesian Approach,” *Journal of Applied Econometrics*, 108-128.
-  Moon, H.R. and F. Schorfheide (2012). “Bayesian and Frequentist Inference in Partially Identified Models” *Econometrica*, 755-782,
-  Chan, J. and Tobias, J.L. (2012). “Priors and Posterior Computation in Linear Endogenous Variables Models with Imperfect Instruments” working paper.

## Conclusion

- MCMC is a very powerful tool that facilitates posterior calculation, and in some cases, enables researchers to estimate models that might otherwise be intractable.
- Here, we illustrated its use in a standard selection or endogenous variable model. Techniques discussed there for handling non-normality or nonlinearities can be adapted to all kinds of other microeconomic models relevant to applied work in Environmental / Resource Economics.
- Remember, we use priors all the time. Bayesian just confess to this practice and incorporate prior information in a way that conforms to the laws of probability theory.
- Thank you so much for allowing me this opportunity.

And, here are a few all-purpose Bayesian textbooks covering a variety of different topics:

-  Poirier, D. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press.
-  Koop, G. (2003). *Bayesian Econometrics*, John Wiley and Sons.
-  Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing.
-  Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, New York: Wiley.
-  Koop, G., D. Poirier and J.L. Tobias (2007). *Bayesian Econometric Methods*, Cambridge: Cambridge University Press.
-  Greenberg, E. (2008). *Introduction to Bayesian Econometrics* Cambridge University Press.