Bayesian McMC data augmentation

outline

1. the Gibbs sampler
2. *McMC* convergence in distribution
3. Albert and Chib's *McMC* data augmented Gibbs probit
4. Metropolis-Hastings bivariate logit (without data augmentation)
5. Li, Poirier, and Tobias' *McMC* data augmented Gibbs treatment effect analysis

- two popular *McMC* strategies originally developed in physical statistical mechanics are the Gibbs sampler, and Metropolis-Hastings (*MH*) algorithm.
- the Gibbs sampler is a special case of *MH*

Bayesian McMC data augmentation

the Gibbs sampler

- suppose we cannot derive $p\left(\theta \mid Y\right)$ in closed form (it does not have a standard probability distribution) but we are able to identify the set of conditional posterior distributions.
- we can utilize the set of full conditional posterior distributions to draw dependent samples for parameters of interest via *McMC* simulation.

Bayesian McMC data augmentation

the Gibbs sampler

- for the full set of conditional posterior distributions

$$p\left(\theta_1 \mid \theta_{-1}, Y\right)$$
$$\vdots$$
$$p\left(\theta_k \mid \theta_{-k}, Y\right)$$

  draws are made for θ_1 conditional on starting values for parameters other than θ_1, that is θ_{-1}.
- then, θ_2 is drawn conditional on the θ_1 draw and the starting values for the remaining θ.
- next, θ_3 is drawn conditional on the draws for θ_1 and θ_2 and the starting values for the remaining θ.
- this continues until all θ have been sampled.
- then the sampling is repeated for a large number of draws with parameters updated each iteration by the most recent draw.

- for example, the procedure for a Gibbs sampler involving two parameters is

1. select a starting value for θ_2,
2. draw θ_1 from $p\,(\theta_1|\theta_2, y)$ utilizing the starting value for θ_2,
3. draw θ_2 from $p(\theta_2|\theta_1, y)$ utilizing the previous draw for θ_1,
4. repeat until a converged sample based on the marginal posteriors is obtained.

Bayesian McMC data augmentation

the Gibbs sampler

- the samples are dependent.
- not all samples will be from the posterior; only after a finite (but unknown) number of iterations are draws from the marginal posterior distribution
- note, in general, $p(\theta_1, \theta_2 \mid Y) \neq p(\theta_1 \mid \theta_2, Y)\, p(\theta_2 \mid \theta_1, Y)$
- convergence is usually checked using trace plots, burn-in iterations, and other convergence diagnostics.
- model specification includes convergence checks, sensitivity to starting values and possibly prior distribution and likelihood assignments, comparison of draws from the posterior predictive distribution with the observed sample, and various goodness of fit statistics.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- Discrete state spaces
- let $S = \left\{\theta^1, \theta^2, \ldots, \theta^d\right\}$ be a discrete state space.
- a Markov chain is a sequence of random variables, $\{\theta_1, \theta_2, \ldots, \theta_r, \ldots\}$ given θ_0 generated by the following transition

$$p_{ij} \equiv \Pr\left(\theta_{r+1} = \theta^j \mid \theta_r = \theta^i\right)$$

- the Markov property says that transition to θ_{r+1} only depends on the immediate past history, θ_r, and not all history.
- define a Markov transition matrix, $P = [p_{ij}]$, where the rows denote initial states and the columns denote transition states such that, for example, p_{ii} is the likelihood of beginning in state i and remaining in state i.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- now, relate this Markov chain idea to distributions from which random variables are drawn.
- say, the initial value, θ_0, is drawn from π_0.
- then, the distribution for θ_1 given $\theta_0 \sim \pi_0$ is

$$\pi_{1j} \equiv \Pr\left(\theta_1 = \theta^j\right) = \sum_{i=1}^d \Pr\left(\theta_0 = \theta^i\right) p_{ij} = \sum_{i=1}^d \pi_{0i} p_{ij}, \quad j = 1, 2, \ldots, d$$

- in matrix notation, the above is

$$\pi_1 = \pi_0 P$$

- and after r iterations we have

$$\pi_r = \pi_0 P^r$$

- as the number of iterations increases, we expect the effect of the initial distribution, π_0, dies out so long as the chain does not get trapped.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- Irreducibility and stationarity
- the idea of no absorbing states or states in which the chain gets trapped is called *irreducibility*.
- if $p_{ij} > 0$ (strictly positive) for all i, j, then the chain is *irreducible* and there exists a *stationary* distribution, π, such that

$$\lim_{r \to \infty} \pi_0 P^r = \pi$$

and

$$\pi P = \pi$$

- since the elements are all positive and each row sums to one, the maximum eigenvalue of P^T is one and π is determined by the corresponding eigenvector, call it S_1, and the corresponding row vector from the inverse of the matrix for eigenvectors, S^{-1}.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- by singular value decomposition, $P = S\Lambda S^{-1}$ where S is a matrix of eigenvectors and Λ is a diagonal matrix of corresponding eigenvalues, $\left(P^T\right)^r = S\Lambda^r S^{-1}$ since

$$\begin{aligned}\left(P^T\right)^r &= S\Lambda S^{-1} S\Lambda S^{-1} \cdots S\Lambda S^{-1} \\ &= S\Lambda^r S^{-1}\end{aligned}$$

- this implies the long-run steady-state is determined by the largest eigenvalue and in the direction of the corresponding vector from the inverse of eigenvector matrix (if the remaining λ's < 1 then λ_i^r goes to zero and their corresponding inverse eigenvectors' influence on direction dies out).
- since one is the largest eigenvalue of P^T, after a large number of iterations $\pi_0 P^r$ converges to $1 \times \pi = \pi$.
- hence, after many iterations the Markov chain produces draws from a stationary distribution if the chain is irreducible.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- Time reversibility and stationarity
- an equivalent property, *time reversibility*, is more useful when working with more general state space chains.
- time reversibility says that if we reverse the order of a Markov chain, the resulting chain has the same transition behavior.
- first, we show the reverse chain is Markovian if the forward chain is Markovian, then we relate the forward and reverse chain transition probabilities, and finally, we show that time reversibility implies $\pi_i p_{ij} = \pi_j p_{ji}$ and this implies $\pi P = \pi$ where π is the stationary distribution for the chain.

- the reverse transition probability (by Bayesian "updating") is

$$
\begin{aligned}
p_{ij}^{*} &\equiv \Pr\left(\theta_r = \theta^{j} \mid \theta_{r+1} = \theta^{i_1}, \theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T}\right) \\
&= \frac{\Pr\left(\theta_r = \theta^{j}, \theta_{r+1} = \theta^{i_1}, \theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T}\right)}{\Pr\left(\theta_{r+1} = \theta^{i_1}, \theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T}\right)} \\
&= \frac{\Pr\left(\theta_r = \theta^{j}\right) \Pr\left(\theta_{r+1} = \theta^{i_1} \mid \theta_r = \theta^{j}\right)}{\Pr\left(\theta_{r+1} = \theta^{i_1}\right)} \\
&\quad \times \frac{\Pr\left(\theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T} \mid \theta_r = \theta^{j}, \theta_{r+1} = \theta^{i_1}\right)}{\Pr\left(\theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T} \mid \theta_{r+1} = \theta^{i_1}\right)}
\end{aligned}
$$

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- since the forward chain is Markovian, we can simplify

$$
\begin{aligned}
p^*_{ij} &= \frac{\Pr\left(\theta_r = \theta^j\right)\Pr\left(\theta_{r+1} = \theta^{i_1} \mid \theta_r = \theta^j\right)}{\Pr\left(\theta_{r+1} = \theta^{i_1}\right)} \\
&\quad \times \frac{\Pr\left(\theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T} \mid \theta_{r+1} = \theta^{i_1}\right)}{\Pr\left(\theta_{r+2} = \theta^{i_2}, \ldots, \theta_{r+T} = \theta^{i_T} \mid \theta_{r+1} = \theta^{i_1}\right)} \\
p^*_{ij} &= \frac{\Pr\left(\theta_r = \theta^j\right)\Pr\left(\theta_{r+1} = \theta^{i_1} \mid \theta_r = \theta^j\right)}{\Pr\left(\theta_{r+1} = \theta^{i_1}\right)}
\end{aligned}
$$

- the reverse chain is Markovian.

- let P^* represent the transition matrix for the reverse chain then the above says
$$p_{ij}^* = \frac{\pi_j p_{ji}}{\pi_i}$$
- by definition, time reversibility implies $p_{ij} = p_{ij}^*$. Hence, time reversibility implies
$$\pi_i p_{ij} = \pi_j p_{ji}$$
- time reversibility says the likelihood of transitioning from state i to j is equal to the likelihood of transitioning from j to i.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- the above implies if a chain is reversible with respect to a distribution π then π is the stationary distribution of the chain.
- to see this sum both sides of the above relation over i

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j \times 1, \quad j = 1, 2, \ldots, d$$

- in matrix notation, we have

$$\pi P = \pi$$

π is the stationary distribution of the chain.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- Continuous state spaces
- continuous state spaces are analogous to discrete state spaces but with a few additional technical details.
- transition probabilities are defined in reference to sets rather than the singletons $\left\{\theta^i\right\}$.
- for example, for a set $A \in \Theta$ the chain is defined in terms of the probabilities of the set given the value of the chain on the previous iteration, θ.
- that is, the kernel of the chain, $K(\theta, A)$, is the probability of set A given the chain is at θ where

$$K(\theta, A) = \int_A p(\theta, \phi)\, d\phi$$

$p(\theta, \phi)$ is a density function with given θ and $p(\cdot, \cdot)$ is the transition or generator function of the kernel.

Bayesian McMC data augmentation

time reversibility and convergence in distribution

- an *invariant* or *stationary* distribution with density $\pi(\cdot)$ implies

$$\int_A \pi(\theta)\,d\theta = \int_\theta K(\theta, A)\,\pi(\theta)\,d\theta = \int_\theta \left[\int_A p(\theta,\phi)\,d\phi\right]\pi(\theta)\,d\theta$$

- *time reversibility* in the continuous space case implies

$$\pi(\theta)\,p(\theta,\phi) = \pi(\phi)\,p(\phi,\theta)$$

- and, *irreducibility* in the continuous state case is satisfied for a chain with kernel K with respect to $\pi(\cdot)$ if every set A with positive probability π can be reached with positive probability after a finite number of iterations.
- in other words, if $\int_A \pi(\theta)\,d\theta > 0$ then there exists $n \geq 1$ such that $K^n(\theta, A) > 0$.
- for continuous state spaces, irreducibility and time reversibility produce a stationary distribution of the chain as with discrete state spaces.

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit

- the challenge with discrete choice models (like probit) is latent utility — the analyst observes only discrete (often binary) choices.
- Albert & Chib [1993] employ Bayesian data augmentation to "supply" the latent variable
- hence, parameters of a probit model are estimated via normal Bayesian regression.
- consider the latent utility model

$$U_D = W\theta - V_D$$

where binary choice, D, is observed

$$D = \begin{cases} 1 & U_D > 0 \\ 0 & U_D < 0 \end{cases}$$

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit

- the conditional posterior distribution for θ is

$$p\left(\theta | D, W, U_D\right) \sim N\left(b_1, \left(Q^{-1} + W^T W\right)^{-1}\right)$$

where

$$b_1 = \left(Q^{-1} + W^T W\right)^{-1} \left(Q^{-1} b_0 + W^T W b\right)$$

$$b = \left(W^T W\right)^{-1} W^T U_D$$

b_0 = prior means for θ and $Q = \left(W_0^T W_0\right)^{-1}$ is the prior for the covariance.

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit

- the conditional posterior distribution for the latent variables are

$$p\left(U_D \middle| D=1, W, \theta\right) \sim N\left(W\theta, I \middle| U_D > 0\right) \text{ or } TN_{(0,\infty)}\left(W\theta, I\right)$$

$$p\left(U_D \middle| D=0, W, \theta\right) \sim N\left(W\theta, I \middle| U_D \leq 0\right) \text{ or } TN_{(-\infty,0)}\left(W\theta, I\right)$$

where $TN(\cdot)$ refers to random draws from a truncated normal (truncated below for the first and truncated above for the second).

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit

- iterative draws for $(U_D|D, W, \theta)$ and $(\theta|D, W, U_D)$ form the Gibbs sampler.
- interval estimates of θ are supplied by post-convergence draws of $(\theta|D, W, U_D)$.
- for simulated normal draws of the unobservable portion of utility, V_D, this Bayesian data augmented probit typically produces similar inferences to *MLE*.

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit prototypical example

- suppose the *DGP* is

$$U_D = -1 + x_1 + x_2 - V_D \quad V_D \sim N(0,1)$$
$$D = \begin{cases} 1 & U_D > 0 \\ 0 & U_D < 0 \end{cases}$$

where x_1 and x_2 are uniform$(0,1)$; in other words, $E[x_1] = E[x_1] = 0.5$

- create a sample of $1,000$ observations
- generate $5,000$ *McMC* data augmented probit draws
- compare *McMC* parameter inference with *MLE*

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit prototypical example

- *MLE* results

$$\Pr\left(D=1 \mid X\right)=\Phi\left(\underset{(0.1142)}{-0.9779}+\underset{(0.1444)}{0.9515}x_1+\underset{(0.1438)}{0.9727}x_2\right)$$

- standard errors are reported in parentheses below parameter estimates
- results are certainly within sampling error of *DGP*

$$\Pr\left(D=1 \mid X\right)=\Phi\left(-1+1x_1+1x_2\right)$$

Bayesian McMC data augmentation

Albert & Chib's data augmented Gibbs sampler probit prototypical example

- *McMC* posterior statistics

statistic	θ_0	θ_1	θ_2
mean	-0.9783	0.9513	0.9740
median	-0.9808	0.9546	0.9747
stand. dev.	0.1136	0.1429	0.1443
0.025 quantile	-1.1946	0.6648	0.6904
0.975 quantile	-0.7526	1.2203	1.2529

- results are remarkably similar to those for *MLE*

Bayesian McMC without data augmentation

Random walk M-H logit prototypical example

The random walk *MH* algorithm employs a standard binary discrete choice model

$$(D_i \mid Z_i) \sim Bernoulli\left(\frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]}\right)$$

The default tuning parameter, $s^2 = 0.25$, produces an apparently satisfactory *MH* acceptance rate of 28.6%. Details are below.

Bayesian McMC without data augmentation

Random walk M-H logit prototypical example

We wish to draw from the posterior

$$\Pr\left(\theta \mid D, Z\right) \propto p\left(\theta\right) \ell\left(\theta \mid D, Z\right)$$

where the log likelihood is

$$\ell\left(\theta \mid D, Z\right) = \sum_{i=1}^{n} D_i \log \frac{\exp\left[Z_i^T \theta\right]}{1 + \exp\left[Z_i^T \theta\right]} + \left(1 - D_i\right) \log\left(1 - \frac{\exp\left[Z_i^T \theta\right]}{1 + \exp\left[Z_i^T \theta\right]}\right)$$

For Z other than a constant, there is no prior, $p\left(\theta\right)$, which produces a well known posterior, $\Pr\left(\theta \mid D, Z\right)$, for the logit model. This makes the *MH* algorithm attractive.

Bayesian McMC without data augmentation

Random walk M-H logit prototypical example

The *MH* algorithm builds a Markov chain (the current draw depends on only the previous draw) such that eventually the influence of initial values dies out and draws are from a stable, approximately independent distribution. The *MH* algorithm applied to the logit model is as follows.

1. Initialize the vector θ^0 at some value.
2. Define a proposal generating density, $q\left(\theta^*, \theta^{k-1}\right)$ for draw $k \in \{1, 2, \ldots, K\}$. The random walk *MH* chooses a convenient generating density.

   $$\theta^* = \theta^{k-1} + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 I\right)$$

   In other words, for each parameter, θ_j,

   $$q\left(\theta_j^*, \theta_j^{k-1}\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{\left(\theta_j^* - \theta_j^{k-1}\right)^2}{2\sigma^2}\right]$$

3. Draw a vector, θ^* from $N\left(\theta^{k-1}, \sigma^2 I\right)$. Notice, for the random walk, the tuning parameter, σ^2, is the key. If σ^2 is chosen too large, then the algorithm will reject the proposal draw frequently and will converge slowly, If σ^2 is chosen too small, then the algorithm will accept the proposal draw frequently but may fail to fully explore the parameter space and may fail to discover the convergent distribution.

4. Calculate $\alpha =$
$$\begin{cases} \min\left(1, \frac{\Pr(\theta^*|D,Z)q\left(\theta^*,\theta^{k-1}\right)}{\Pr\left(\theta^{k-1}|D,Z\right)q\left(\theta^{k-1},\theta^*\right)}\right) & \Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right) > 0 \\ 1 & \Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right) = 0 \end{cases}$$
The core of the *MH* algorithm is that the ratio eliminates the problematic normalizing constant for the posterior (normalization is problematic since we don't recognize the posterior). The convenience of the random walk MH enters here as, by symmetry of the normal, $q\left(\theta^*, \theta^{k-1}\right) = q\left(\theta^{k-1}, \theta^*\right)$ and the calculation of α simplifies as $\frac{q\left(\theta^*,\theta^{k-1}\right)}{q\left(\theta^{k-1},\theta^*\right)}$ drops out. Hence, we calculate
$$\alpha = \begin{cases} \min\left(1, \frac{\Pr(\theta^*|D,Z)}{\Pr\left(\theta^{k-1}|D,Z\right)}\right) & \Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right) > 0 \\ 1 & \Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right) = 0 \end{cases}$$

Bayesian McMC without data augmentation

Random walk M-H logit prototypical example

5. Draw U from a Uniform$(0,1)$. If $U < \alpha$, set $\theta^k = \theta^*$, otherwise set $\theta^k = \theta^{k-1}$. In other words, with probability α accept the proposal draw, θ^*.
6. Repeat K times until the distribution converges.

Bayesian McMC without data augmentation

Random walk M-H logit prototypical example

ML logit estimates (with standard errors in parentheses below the estimates) are

$$E\left[U_D \mid Z\right] = -\underset{(0.3514)}{0.9500}Z_1 + \underset{(0.2419)}{0.7808}Z_2 - \underset{(0.4209)}{0.2729}Z_3 - \underset{(0.3250)}{1.1193}Z_4 + \underset{(0.3032)}{0.3385}Z_5$$

Logit results are proportional to the probit results (approximately 1.5 times the probit estimates), as is typical. As with the probit model, the logit model has modest explanatory power (pseudo-$R^2 = 1 - \frac{\ell(Z\widehat{\theta})}{\ell(\widehat{\theta}_0)} = 10.8\%$, where $\ell\left(Z\widehat{\theta}\right)$ is the log-likelihood for the model and $\ell\left(\widehat{\theta}_0\right)$ is the log-likelihood with a constant only).

Now, we compare the *ML* results with *McMC* posterior draws. Statistics from $10,000$ posterior *MH* draws following $1,000$ burn-in draws are tabulated below based on the $n = 120$ sample.

statistic	θ_1	θ_2	θ_3	θ_4	θ_5
mean	-0.9850	0.8176	-0.2730	-1.1633	0.3631
median	-0.9745	0.8066	-0.2883	-1.1549	0.3440
standard deviation	0.3547	0.2426	0.4089	0.3224	0.3069
quantiles:					
0.025	-1.7074	0.3652	-1.0921	-1.7890	-0.1787
0.25	-1.2172	0.6546	-0.5526	-1.3793	0.1425
0.75	-0.7406	0.9787	0.0082	-0.9482	0.5644
0.975	-0.3134	1.3203	0.5339	-0.5465	0.9924

Sample statistics for *MH McMC* logit posterior draws

$$DGP : U_D = Z\theta + \varepsilon, \quad Z = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 \end{bmatrix}$$

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects

- the challenge with treatment effects is latent utility and counterfactuals
- following Albert & Chib [1993] we employ Bayesian data augmentation to "supply" the latent variable and counterfactuals.
- hence, parameters of the selection model are estimated via Bayesian seemingly unrelated regression.
- another challenge is $Corr\left(V_1, V_0\right)$ is unidentified (due to counterfactuals), LPT utilize bounding to address this issue.

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects—bounds and learning

- Even if we know Δ is normally distributed, unobservability of the counterfactuals creates a problem for identifying the distribution of Δ as
$$Var\left[\Delta \mid X\right] = Var\left[V_1\right] + Var\left[V_0\right] - 2Cov\left[V_1, V_0\right]$$
and $\rho_{10} \equiv Corr\left[V_1, V_0\right]$ is unidentified.
- Let $\eta \equiv \left[V_D, V_1, V_0\right]^T$ then
$$\Sigma \equiv Var\left[\eta\right] = \begin{bmatrix} 1 & \rho_{D1}\sigma_1 & \rho_{D0}\sigma_0 \\ \rho_{D1}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{D0}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{bmatrix}$$

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects—bounds and learning

- From the positivity of the determinant (or eigenvalues) of Σ we can bound the unidentified correlation

$$\rho_{10} \in \rho_{D1}\rho_{D0} \pm \left[\left(1-\rho_{D1}^{2}\right)\left(1-\rho_{D0}^{2}\right)\right]^{\frac{1}{2}}$$

- This allows learning about ρ_{10} and, in turn, identification of the distribution of treatment effects.
- Notice the more pressing is the endogeneity problem (ρ_{D1}, ρ_{D0} large in absolute value) the tighter are the bounds.

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects

- consider the latent utility model

$$D^* = -1 + x + z + V_D$$

where binary choice, D, is observed

$$D = \begin{cases} 1 & D^* > 0 \\ 0 & D^* < 0 \end{cases}$$

- and outcome equations

$$\begin{aligned} Y_1 &= 2 + 10x + V_1 \\ Y_0 &= 1 + 2x + V_0 \end{aligned}$$

$E[x] = E[z] = 0.5$

- with variance-covariance for $[V_D, V_1, V_0]$

$$\Sigma = \begin{bmatrix} 1 & 0.7 & -0.7 \\ 0.7 & 1 & -0.1 \\ -0.7 & -0.1 & 1 \end{bmatrix}$$

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects

- average treatment effects

$$ATE = (2-1) + (10-2)\,0.5 = 5$$

$$ATT = 5 + [0.7 - (-0.7)]\,(0.8) = 6.12$$

$$ATUT = 5 + [0.7 - (-0.7)]\,(-0.8) = 3.88$$

where $E\left[\frac{\phi(W\theta)}{\Phi(W\theta)} \mid D=1\right] \approx 0.8$ and $E\left[-\frac{\phi(W\theta)}{\Phi(-W\theta)} \mid D=0\right] \approx -0.8$

- outcome is heterogeneous
- effect identified by *OLS*

$$OLS = 2 + 10\,(0.5) - [1 + 2\,(0.5)] = 5$$

- common mistake is to limit comparison of *ATE* with *OLS* to assess impact of endogeneity

- define the complete or augmented data as

$$r_i^* = \begin{bmatrix} D_i^* & D_i Y_i + (1 - D_i)\, Y_i^{miss} & D_i Y_i^{miss} + (1 - D_i)\, Y_i \end{bmatrix}^T$$

- also, let

$$H_i = \begin{bmatrix} W_i & 0 & 0 \\ 0 & X_i & 0 \\ 0 & 0 & X_i \end{bmatrix}$$

and

$$\beta = \begin{bmatrix} \theta \\ \beta_1 \\ \beta_0 \end{bmatrix}$$

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects

- full conditional posterior distributions
- let Γ_{-x} denote all parameters other than x.
- the full conditional posteriors for the augmented outcome data are

$$Y_i^{miss} \mid \Gamma_{-Y_i^{miss}}, Data \sim N\left(\left(1-D_i\right)\mu_{1i} + D_i\mu_{0i}, \left(1-D_i\right)\omega_{1i} + D_i\omega_{0i}\right)$$

where standard multivariate normal theory is applied to derive means and variances conditional on the draw for latent utility and the other outcome

$$\mu_{1i} = X_i\beta_1 + \frac{\sigma_0^2\sigma_{D1} - \sigma_{10}\sigma_{D0}}{\sigma_0^2 - \sigma_{D0}^2}\left(D_i^* - Z_i\theta\right) + \frac{\sigma_{10} - \sigma_{D1}\sigma_{D0}}{\sigma_0^2 - \sigma_{D0}^2}\left(Y_i - X_i\beta_0\right)$$

$$\mu_{0i} = X_i\beta_0 + \frac{\sigma_1^2\sigma_{D0} - \sigma_{10}\sigma_{D1}}{\sigma_1^2 - \sigma_{D1}^2}\left(D_i^* - Z_i\theta\right) + \frac{\sigma_{10} - \sigma_{D1}\sigma_{D0}}{\sigma_1^2 - \sigma_{D1}^2}\left(Y_i - X_i\beta_1\right)$$

$$\omega_{1i} = \sigma_1^2 - \frac{\sigma_{D1}^2 \sigma_0^2 - 2\sigma_{10}\sigma_{D1}\sigma_{D0} + \sigma_{10}^2}{\sigma_0^2 - \sigma_{D0}^2}$$

$$\omega_{0i} = \sigma_0^2 - \frac{\sigma_{D0}^2 \sigma_1^2 - 2\sigma_{10}\sigma_{D1}\sigma_{D0} + \sigma_{10}^2}{\sigma_1^2 - \sigma_{D1}^2}$$

- similarly, the conditional posterior for the latent utility is

$$D_i^* \mid \Gamma_{-D_i^*}, Data \sim \begin{array}{cc} TN_{(0,\infty)}\left(\mu_{D_i}\omega_D\right) & if \;\; D_i = 1 \\ TN_{(-\infty,0)}\left(\mu_{D_i}\omega_D\right) & if \;\; D_i = 0 \end{array}$$

where $TN(\cdot)$ refers to the truncated normal distribution with support indicated via the subscript and the arguments are parameters of the untruncated distribution.

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects

- applying multivariate normal theory for $(D_i^* \mid Y_i)$ we have

$$\begin{aligned}
\mu_{D_i} &= Z_i\theta + \left(D_i Y_i + (1 - D_i)\, Y_i^{miss} - X_i\beta_1\right) \frac{\sigma_0^2 \sigma_{D1} - \sigma_{10}\sigma_{D0}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2} \\
&\quad + \left(D_i Y_i^{miss} + (1 - D_i)\, Y_i - X_i\beta_0\right) \frac{\sigma_1^2 \sigma_{D0} - \sigma_{10}\sigma_{D1}}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}
\end{aligned}$$

$$\omega_D = 1 - \frac{\sigma_{D1}^2\sigma_0^2 - 2\sigma_{10}\sigma_{D1}\sigma_{D0} + \sigma_{D0}^2\sigma_1^2}{\sigma_1^2\sigma_0^2 - \sigma_{10}^2}$$

- the conditional posterior distribution for the parameters is

$$\beta \mid \Gamma_{-\beta}, Data \sim N\left(\mu_\beta, \omega_\beta\right)$$

where by the *SUR* (seemingly-unrelated regression) generalization of Bayesian regression

$$\mu_\beta = \left[H^T\left(\Sigma^{-1} \otimes I_n\right) H + V_\beta^{-1}\right]^{-1} \left[H^T\left(\Sigma^{-1} \otimes I_n\right) r^* + V_\beta^{-1} \beta_0\right]$$

$$\omega_\beta = \left[H^T\left(\Sigma^{-1} \otimes I_n\right) H + V_\beta^{-1}\right]^{-1}$$

and the prior distribution is $p(\beta) \sim N\left(\beta_0, V_\beta\right)$.

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented Gibbs sampler for treatment effects

- the conditional distribution for the trivariate variance-covariance matrix is

$$\Sigma \mid \Gamma_{-\Sigma}, Data \sim G^{-1}$$

where

$$G \sim Wishart\left(n+\rho, S+\rho R\right)$$

with prior $p\left(G\right) \sim Wishart\left(\rho, \rho R\right)$, and

$S = \sum_{i=1}^{n} \left(r_i^* - H_i \beta\right)\left(r_i^* - H_i \beta\right)^T$.

- Nobile's algorithm
- recall σ_D^2 is normalized to one; this creates a slight complication as the conditional posterior is no longer inverse-Wishart.
- the algorithm applied to the current setting results in the following steps:

1. Exchange rows and columns one and three in $S + \rho R$, call this matrix V.
2. Find L such that $V = \left(L^{-1}\right)^T L^{-1}$.
3. Construct a lower triangular matrix A with
   a. a_{ii} equal to the square root of χ^2 random variates, $i = 1, 2$.
   b. $a_{33} = \frac{1}{l_{33}}$ where l_{33} is the third row-column element of L.
   c. a_{ij} equal to $N(0, 1)$ random variates, $i > j$.
4. Set $V' = \left(L^{-1}\right)^T \left(A^{-1}\right)^T A^{-1} L^{-1}$.
5. Exchange rows and columns one and three in V' and denote this draw Σ.

- prior distributions
- Li, Poirier, and Tobias choose relatively diffuse priors such that the data dominates the posterior distribution.
- their prior distribution for β is $p(\beta) \sim N(\beta_0, V_\beta)$ where $\beta_0 = 0$, $V_\beta = 4I$ and their prior for Σ^{-1} is $p(G) \sim Wishart(\rho, \rho R)$ where $\rho = 12$ and R is a diagonal matrix with elements $\left\{\frac{1}{12}, \frac{1}{4}, \frac{1}{4}\right\}$.

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented treatment effect example

- create a sample of $1,000$ observations
- generate $5,000$ *McMC* data augmented probit draws
- compare *McMC* parameter inference with Heckman's two-stage inverse Mills strategy and sample average treatment effects

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented treatment effect example

- results
- sample statistics based on "known" counterfactuals

$$
\begin{aligned}
ATE &= \overline{Y_1} - \overline{Y_0} = 5.0545 \\
ATT &= \frac{\sum D_i \left(Y_{1i} - Y_{0i} \right)}{\sum D_i} = 6.6528 \\
ATUT &= \frac{\sum \left(1 - D_i \right) \left(Y_{1i} - Y_{0i} \right)}{\sum \left(1 - D_i \right)} = 3.4561 \\
OLS &= \frac{\sum D_i Y_{1i}}{\sum D_i} - \frac{\sum \left(1 - D_i \right) Y_{0i}}{\sum \left(1 - D_i \right)} = 5.8444
\end{aligned}
$$

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented treatment effect example

- results — Heckman's two stage
- selection equation

$$\begin{aligned} \Pr(D \mid W) &= \Phi(-0.9779 + 0.9515x + 0.9727z) \\ pseudo - R^2 &= 0.0627 \end{aligned}$$

- outcomes

$$\begin{aligned} E[Y \mid D, W] &= 1.3319 + 4.7493D + 2.0094x \\ &\quad +7.9023D(x - \overline{x}) + 0.7261D\frac{\phi\left(W\widehat{\theta}\right)}{\Phi\left(W\widehat{\theta}\right)} \\ &\quad -0.3470(1 - D)\frac{-\phi\left(W\widehat{\theta}\right)}{\Phi\left(-W\widehat{\theta}\right)} \end{aligned}$$

- results
- sample statistics from Heckman two stage regression

$$\begin{aligned} estATE &= 4.7493 \\ estATT &= 5.5400 \\ estATUT &= 3.9598 \end{aligned}$$

- *McMC* results

statistic	θ_0	θ_1	θ_2
mean	-1.0999	1.1151	1.0420
median	-1.1027	1.1152	1.0405
stand. dev.	0.1155	0.1487	0.1420
0.025 quantile	-1.3211	0.8219	0.7653
0.975 quantile	-0.8749	1.3961	1.3149

- *McMC* results for

$$E\left[Y \mid D, W\right] = \beta_{00}\left(1 - D\right) + \beta_{01}D + \beta_{10}\left(1 - D\right)x + \beta_{11}Dx$$

statistic	β_{00}	β_{10}	β_{10}	β_{11}
mean	1.4037	2.0471	2.4090	9.7316
median	1.4018	2.0461	2.4038	9.7334
std. dev.	0.1135	0.1880	0.1677	0.1610
quantile				
0.025	1.1847	1.6761	2.0980	9.4189
0.975	1.6258	2.4110	2.7578	10.0373

- *McMC* results

statistic	$\rho_{D,1}$	$\rho_{D,0}$	$\rho_{1,0}$
mean	0.3847	-0.2413	-0.1297
median	0.4033	-0.2468	-0.1284
stand. dev.	0.1361	0.1947	0.1863
0.025 quantile	0.0772	-0.5857	-0.4804
0.975 quantile	0.6096	0.1494	0.2170

- *McMC* results

statistic	*ATE*	*ATT*	*ATUT*
mean	4.9596	5.8979	4.0212
median	4.9640	5.8951	4.0063
stand. dev.	0.1681	0.2981	0.2058
0.025 quantile	4.6402	5.3285	3.6520
0.975 quantile	5.2869	6.4820	4.4525
sample stat.	5.0545	6.6528	3.4561
Heckman 2SLS	4.7493	5.5400	3.9598

Bayesian McMC data augmentation

Li, Poirier & Tobias' data augmented treatment effect example

- in this case, *McMC* results are closer to the sample statistics than is the Heckman two stage strategy
- more work to do including
- Bayesian identification of marginal treatment effects and connections to average treatment effects
- further incorporation of background knowledge in the Bayesian strategy
- Metropolis-Hastings (*McMC*) strategies when full set of conditional posteriors is not identified