# Contents

# 7
# Bayesian simulation

In the chapter, we discuss direct and conditional posterior simulation as well as Markov chain Monte Carlo (*McMC*) simulation.

## 7.1 Direct posterior simulation

Posterior simulation allows us to learn about features of the posterior (including linear combinations, products, or ratios of parameters) by drawing samples when the exact form of the posterior density is analytically intractable. Monte Carlo simulation implies we employ a sufficiently large number of draws to (approximately) cover the sample space for the quantities of interest.

*Example*

For example, suppose $x_1$ and $x_2$ are (jointly) Gaussian or normally distributed with unknown means, $\begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}$, and known variance-covariance, $\sigma^2 I_2 = 9I$, but we're interested in $x_3 = \sqrt{\alpha_1 x_1 + \alpha_2 x_2}$. Based on a sample of data ($n = 30$), $y = \{x_1, x_2\}$, we can infer the posterior means and variance for $x_1$ and $x_2$ and simulate posterior draws for $x_3$ from which properties of the posterior distribution for $x_3$ can be inferred. Suppose $\mu_1 = 50$ and $\mu_2 = 75$ and we have no prior knowledge regarding the location of $x_1$ and $x_2$ so we employ uniform (uninformative) priors. Sample statistics for

$x_1$ and $x_2$ are reported below.

| statistic | $x_1$ | $x_2$ |
|---|---|---|
| mean | 51.0 | 75.5 |
| median | 50.8 | 76.1 |
| standard deviation | 3.00 | 2.59 |
| maximum | 55.3 | 80.6 |
| minimum | 43.6 | 69.5 |
| quantiles: | | |
| 0.01 | 43.8 | 69.8 |
| 0.025 | 44.1 | 70.3 |
| 0.05 | 45.6 | 71.1 |
| 0.10 | 47.8 | 72.9 |
| 0.25 | 49.5 | 73.4 |
| 0.75 | 53.0 | 77.4 |
| 0.9 | 54.4 | 78.1 |
| 0.95 | 55.1 | 79.4 |
| 0.975 | 55.3 | 80.2 |
| 0.99 | 55.3 | 80.5 |

Sample statistics

Since we know $x_1$ and $x_2$ are independent each with variance 9, the marginal posteriors for $\mu_1$ and $\mu_2$ are

$$p\left(\mu_1 \mid y\right) \sim N\left(\overline{x_1} = 51.0, \frac{9}{30}\right)$$

and

$$p\left(\mu_2 \mid y\right) \sim N\left(\overline{x_2} = 75.5, \frac{9}{30}\right)$$

and the predictive posteriors for $x_1$ and $x_2$ are based on posteriors draws for $\mu_1$ and $\mu_2$

$$p\left(x_1 \mid \mu_1, y\right) \sim N\left(\mu_1, 9\right)$$

and

$$p\left(x_2 \mid \mu_2, y\right) \sim N\left(\mu_2, 9\right)$$

The posterior distributions for the means are proportional to their likelihood functions.

$$\ell\left(\mu_j; x_j, \sigma\right) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}\left(x_{ji} - \mu\right)^2\right]$$

This expands (by completing the square) and simplifies.

$$\ell\left(\mu_j; x_j, \sigma\right) \propto \exp\left[-\frac{n}{2\sigma^2}\left(\overline{x_j} - \mu\right)^2\right] \exp\left[-\frac{n}{2\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ji}^2 - \overline{x_j}^2\right)^2\right]$$

The second term does not involve $\mu_j$ so it is absorbed into the normalizing constant and the likelihood simplifies as

$$\ell\left(\mu_j; x_j, \sigma\right) \propto \exp\left[-\frac{1}{2\sigma^2/n}\left(\overline{x_j} - \mu\right)^2\right]$$

In other words, as indicated above, $\mu_j$ is drawn from a normal distribution with mean $\overline{x_j}$ and variance $\frac{\sigma^2}{n}$.

For $\alpha_1 = 2$ and $\alpha_2 = 3$, we generate $1,000$ posterior predictive draws of $x_1$ and $x_2$, and utilize them to create posterior predictive draws for $x_3$. Sample statistics for these posterior draws are reported below.

| statistic | $\mu_1$ | $\mu_1$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|---|
| mean | 51.0 | 75.5 | 50.9 | 75.4 | 149.2 |
| median | 51.0 | 75.5 | 50.8 | 75.3 | 149.2 |
| standard deviation | 0.55 | 0.56 | 3.15 | 3.04 | 5.07 |
| maximum | 52.5 | 77.5 | 59.7 | 85.4 | 163.1 |
| minimum | 48.5 | 73.5 | 39.4 | 65.4 | 131.4 |
| quantiles: | | | | | |
| 0.01 | 49.6 | 74.4 | 44.1 | 68.5 | 137.8 |
| 0.025 | 49.8 | 74.5 | 44.7 | 69.7 | 139.8 |
| 0.05 | 50.1 | 74.6 | 45.7 | 70.6 | 141.2 |
| 0.10 | 50.3 | 74.8 | 46.8 | 71.6 | 142.8 |
| 0.25 | 50.6 | 75.2 | 48.8 | 73.3 | 145.7 |
| 0.75 | 51.3 | 75.9 | 52.9 | 77.6 | 152.8 |
| 0.9 | 51.6 | 76.3 | 55.0 | 79.4 | 155.6 |
| 0.95 | 51.8 | 76.5 | 56.2 | 80.5 | 157.6 |
| 0.975 | 52.0 | 76.7 | 57.5 | 81.6 | 159.4 |
| 0.99 | 52.3 | 76.9 | 58.5 | 82.4 | 160.9 |

Sample statistics for posterior draws

A normal probability plot[1] and histogram based on $1,000$ draws of $x_3$ along with the descriptive statistics above based on posterior draws suggest

[1] We employ Filliben's [1975] approach by plotting normal quantiles of $u_j$, $N(u_i)$, (horizontal axis) against $z$ scores (vertical axis) for the data, $y$, of sample size $n$ where

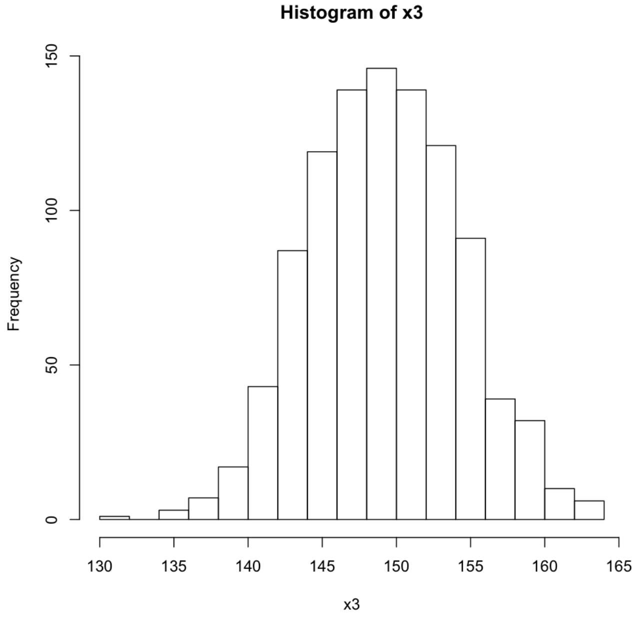
$$u_i = \begin{array}{cc} 1 - 0.5^n & j = 1 \\ \frac{j-0.3175}{n+0.365} & j = 2, \ldots, n-1 \\ 0.5^n & j = n \end{array}$$

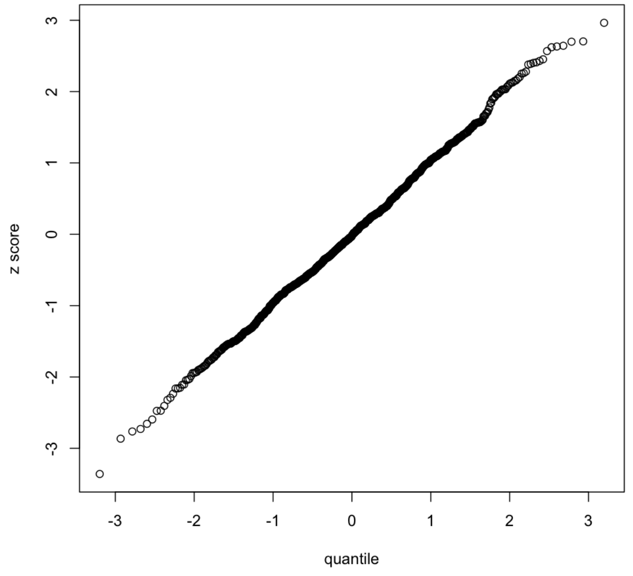
(a general expression is $\frac{j-a}{n+1-2a}$, in the above $a = 0.3175$), and

$$z_i = \frac{y_i - \overline{y}}{s}$$

with sample average, $\overline{y}$, and sample standard deviation, $s$.

that $x_3$ is well approximated by a Gaussian distribution.





Normal probability plot for $x_3$

## 7.2 Independent simulation

The above example illustrates independent simulation. Since $x_1$ and $x_2$ are independent, their joint distribution, $p(x_1, x_2)$, is the product of their marginals, $p(x_1)$ and $p(x_2)$. As these marginals depend on their unknown means, we can independently draw from the marginal posteriors for the means, $p(\theta_1 \mid y)$ and $p(\theta_2 \mid y)$, to generate predictive posterior draws for $x_1$ and $x_2$.[2]

The general independent posterior simulation procedure is

1. draw $\theta_2$ from the marginal posterior $p(\theta_2 \mid y)$,
2. draw $\theta_1$ from the marginal posterior $p(\theta_1 \mid y)$.

## 7.3 Nuisance parameters & conditional simulation

When there are nuisance parameters or, in other words, the model is hierarchical in nature, it is simpler to employ conditional posterior simulation. That is, draw the nuisance parameter from its marginal posterior then draw the other parameters of interest conditional on the draw of the nuisance or hierarchical parameter.

The general conditional simulation procedure is

1. draw $\theta_2$ (say, scale) from the marginal posterior $p(\theta_2 \mid y)$,
2. draw $\theta_1$ (say, mean) from the conditional posterior $p(\theta_1 \mid \theta_2, y)$.

Again, a key is we employ a sufficiently large number of draws to "fully" sample $\theta_1$ conditional on "all" $\theta_2$.

*Example*

We compare independent simulation based on marginal posteriors for the mean and variance with conditional simulation based on the marginal posterior of the variance and the conditional posterior of the mean for the Gaussian (normal) unknown mean and variance case. First ,we explore informed priors, then we compare with uninformative priors. An exchangeable sample of $n = 50$ observations from a Gaussian (normal) distribution with mean equal to 46, a draw from the prior distribution for the mean (described below), and variance equal to 9, a draw from the prior distribution for the variance (also, described below).

[2] Predictive posterior simulation is discussed below where we add another simulation step involving a predictive draw conditional on the observed data and simulated parameters.

### *7.3.1 Informed priors*

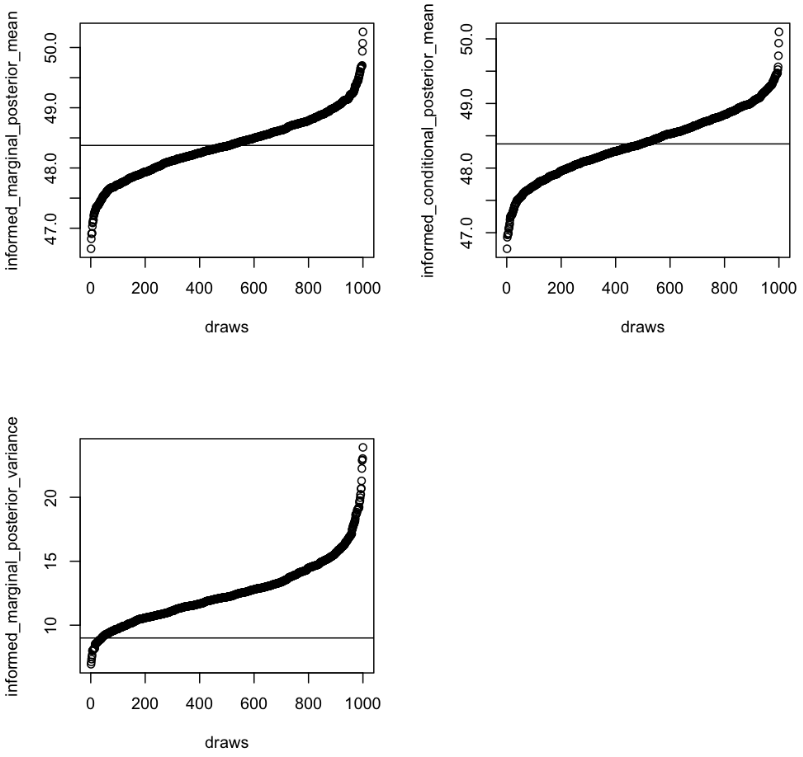
The prior distribution for the mean is Gaussian with mean equal to $\theta_0 = 50$ and variance equal to $\frac{\sigma^2}{\kappa_0} = 18$ ($\kappa_0 = \frac{1}{2}$). The prior distribution for the variance is inverted-chi square with $\nu_0 = 5$ degrees of freedom and scale equal to $\sigma_0^2 = 9$. Then, the marginal posterior for the variance is inverted-chi square with $\nu_n = \nu_0 + n = 55$ degrees of freedom and scale equal to $\nu_n \sigma_n^2 = 45 + 49s^2 + \frac{25}{50.5}\left(50 - \overline{y}\right)^2$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ depend on the sample. The conditional posterior for the mean is Gaussian with mean equal to $\theta_n = \frac{1}{50.5}\left(25 + 50\overline{y}\right)$ and variance equal to the draw from marginal posterior for the variance scaled by $\kappa_0 + n$, $\frac{\sigma^2}{50.5}$. The marginal posterior for the mean is noncentral, scaled Student t with noncentrality parameter equal to $\theta_n = \frac{1}{50.5}\left(25 + 50\overline{y}\right)$ and scale equal to $\frac{\sigma_n^2}{50.5}$. In other words, posterior draws for the mean are $\theta = t\sqrt{\frac{\sigma_n^2}{50.5}} + \theta_n$ where $t$ is a draw from a standard Student t(55) distribution.

Statistics for $1,000$ marginal and conditional posterior draws of the mean and marginal posterior draws of the variance are tabulated below.

| statistic | $(\theta \mid y)$ | $(\theta \mid \sigma^2, y)$ | $(\sigma^2 \mid y)$ |
|---|---|---|---|
| mean | 45.4 | 45.5 | 9.6 |
| median | 45.4 | 45.5 | 9.4 |
| standard deviation | 0.45 | 0.44 | 1.85 |
| maximum | 46.8 | 46.9 | 21.1 |
| minimum | 44.1 | 43.9 | 5.5 |
| quantiles: | | | |
| 0.01 | 44.4 | 44.4 | 6.1 |
| 0.025 | 44.5 | 44.6 | 6.6 |
| 0.05 | 44.7 | 44.8 | 7.0 |
| 0.10 | 44.9 | 44.8 | 7.0 |
| 0.25 | 45.1 | 45.2 | 8.3 |
| 0.75 | 45.7 | 45.8 | 10.7 |
| 0.9 | 46.0 | 46.0 | 12.0 |
| 0.95 | 46.2 | 46.2 | 12.8 |
| 0.975 | 46.3 | 46.3 | 13.4 |
| 0.99 | 46.5 | 46.5 | 14.9 |

Sample statistics for posterior draws based on informed priors

Clearly, marginal and conditional posterior draws for the mean are very similar, as expected. Marginal posterior draws for the variance have more spread than those for the mean, as expected, and all posterior draws comport well with the underlying distribution. Sorted posterior draws based on informed priors are plotted below with the underlying parameter depicted by a horizontal line.



Posterior draws for the mean and variance based on informed priors

As the evidence and priors are largely in accord, we might expect the informed priors to reduce the spread in the posterior distributions somewhat. Below we explore uninformed priors.

## *7.3.2 Uninformed priors*

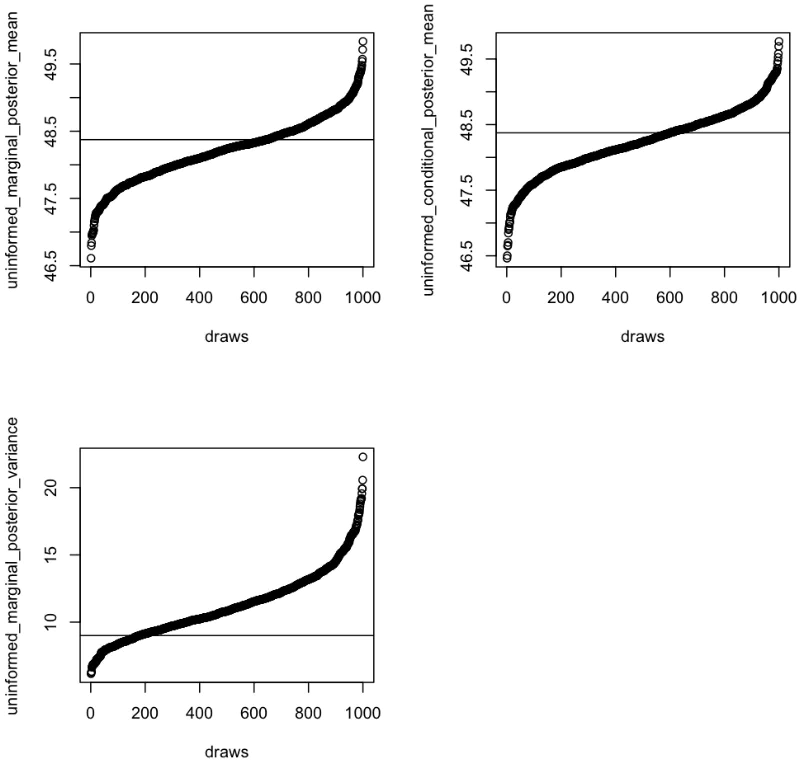
The marginal posterior for the variance is inverted-chi square with $n - 1 = 49$ degrees of freedom and scale equal to $(n-1)\,s^2 = 49s^2$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ depend on the sample. The conditional posterior for the mean is Gaussian with mean equal to $\overline{y}$ and variance equal to the draw from marginal posterior for the variance scaled by $n$, $\frac{\sigma^2}{50}$. The marginal posterior for the mean is noncentral, scaled Student t with noncentrality parameter equal to $\overline{y}$ and scale equal to $\frac{s^2}{50}$. In other words, posterior draws for the mean are $\theta = t\sqrt{\frac{s^2}{50}} + \overline{y}$ where $t$ is a draw from a standard Student t(49) distribution.

Statistics for $1,000$ marginal and conditional posterior draws of the mean and marginal posterior draws of the variance are tabulated below.

| statistic | $(\theta \mid y)$ | $(\theta \mid \sigma^2, y)$ | $(\sigma^2 \mid y)$ |
|---|---|---|---|
| mean | 45.4 | 45.4 | 9.7 |
| median | 45.4 | 45.5 | 9.4 |
| standard deviation | 0.43 | 0.45 | 2.05 |
| maximum | 46.7 | 47.0 | 18.9 |
| minimum | 44.0 | 43.9 | 4.8 |
| quantiles: | | | |
| 0.01 | 44.4 | 44.3 | 6.2 |
| 0.025 | 44.6 | 44.5 | 6.6 |
| 0.05 | 44.7 | 44.7 | 6.9 |
| 0.10 | 44.9 | 44.8 | 7.3 |
| 0.25 | 45.1 | 45.1 | 8.3 |
| 0.75 | 45.7 | 45.7 | 10.9 |
| 0.9 | 46.0 | 46.0 | 12.4 |
| 0.95 | 46.1 | 46.2 | 13.5 |
| 0.975 | 46.3 | 46.3 | 14.3 |
| 0.99 | 46.4 | 46.4 | 15.6 |

Sample statistics for posterior draws based on informed priors

There is remarkably little difference between the informed and uninformed posterior draws. With a smaller sample we would expect the priors to have a more substantial impact. Sorted posterior draws based on uninformed priors are plotted below with the underlying parameter depicted by a horizontal line.



Posterior draws for the mean and variance based on uninformed priors

### *7.3.3 Discrepant evidence*

Before we leave this subsection, perhaps it is instructive to explore the implications of discrepant evidence. That is, we investigate the case where the evidence differs substantially from the priors. We again draw a value for $\theta$ from a Gaussian distribution with mean 50 and variance $\frac{9}{1/2}$, now the draw is $\theta = 53.1$. Then, we set the prior for $\theta$, $\theta_0$, equal to $50+6\frac{\sigma}{\sqrt{\kappa_0}} = 75.5$. Everything else remains as above. As expected, posterior draws based on uninformed priors are very similar to those reported above except with the shift in the mean for $\theta$.[3]

[3] To conserve space, posterior draws based on the uninformed prior results are not reported.

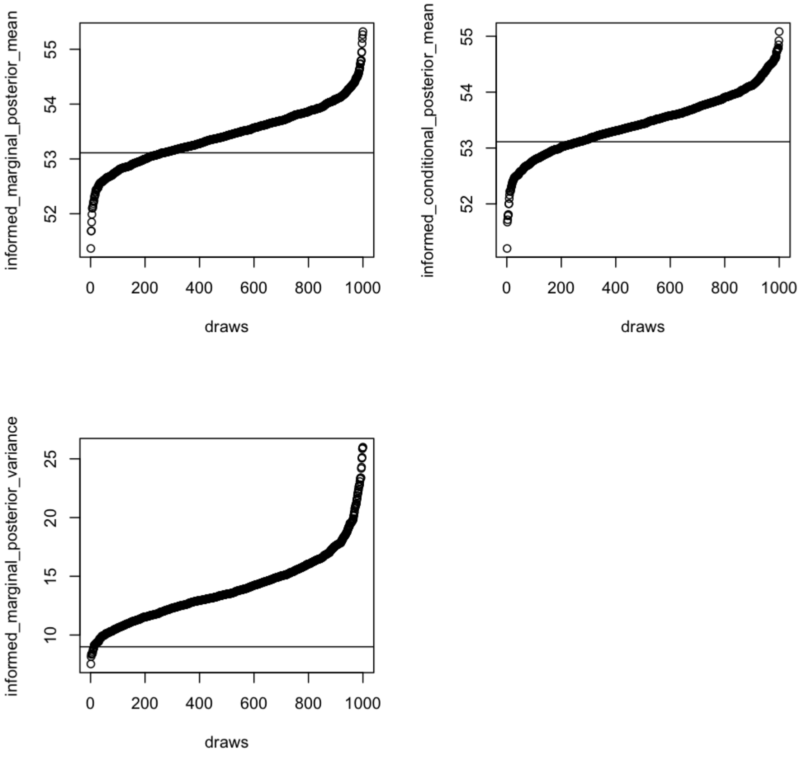
Based on informed priors, the marginal posterior for the variance is inverted-chi square with $\nu_n = \nu_0 + n = 55$ degrees of freedom and scale equal to $\nu_n \sigma_n^2 = 45 + 49s^2 + \frac{25}{50.5}\left(75.5 - \overline{y}\right)^2$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ depend on the sample. The conditional posterior for the mean is Gaussian with mean equal to $\theta_n = \frac{1}{50.5}\left(37.75 + 50\overline{y}\right)$ and variance equal to the draw from marginal posterior for the variance scaled by $\kappa_0 + n$, $\frac{\sigma^2}{50.5}$. The marginal posterior for the mean is noncentral, scaled Student t with noncentrality parameter equal to $\theta_n = \frac{1}{50.5}\left(37.75 + 50\overline{y}\right)$ and scale equal to $\frac{\sigma_n^2}{50.5}$. In other words, posterior draws for the mean are $\theta = t\sqrt{\frac{\sigma_n^2}{50.5}} + \theta_n$ where $t$ is a draw from a standard Student t(55) distribution.

Statistics for $1,000$ marginal and conditional posterior draws of the mean and marginal posterior draws of the variance are tabulated below.

| statistic | $(\theta \mid y)$ | $(\theta \mid \sigma^2, y)$ | $(\sigma^2 \mid y)$ |
|---|---|---|---|
| mean | 53.4 | 53.4 | 13.9 |
| median | 53.4 | 53.4 | 13.5 |
| standard deviation | 0.52 | 0.54 | 2.9 |
| maximum | 55.3 | 55.1 | 26.0 |
| minimum | 51.4 | 51.2 | 7.5 |
| quantiles: | | | |
| 0.01 | 52.2 | 52.2 | 8.8 |
| 0.025 | 52.5 | 52.4 | 9.4 |
| 0.05 | 52.6 | 52.6 | 10.0 |
| 0.10 | 52.8 | 52.8 | 10.6 |
| 0.25 | 53.1 | 53.1 | 11.9 |
| 0.75 | 53.8 | 53.8 | 15.5 |
| 0.9 | 54.1 | 54.1 | 17.6 |
| 0.95 | 54.3 | 54.4 | 19.2 |
| 0.975 | 54.5 | 54.5 | 21.1 |
| 0.99 | 54.7 | 54.7 | 23.3 |

Sample statistics for posterior draws based on informed priors: discrepant case

Posterior draws for $\theta$ are largely unaffected by the discrepancy between the evidence and the prior, presumably, because the evidence dominates with a sample size of 50. However, consistent with intuition posterior draws for the variance are skewed upward more than previously. Sorted posterior draws based on informed priors are plotted below with the underlying parameter depicted by a horizontal line.



Posterior draws for the mean and variance based on informed priors: discrepant case

## 7.4 Posterior predictive distribution

As we've seen for independent simulation (the first example in this section), posterior predictive draws allow us to generate distributions for complex combinations of parameters or random variables.

For independent simulation, the general procedure for generating posterior predictive draws is

1. draw $\theta_1$ from $p\,(\theta_1 \mid y)$,
2. draw $\theta_2$ from $p\,(\theta_2 \mid y)$,
3. draw $\widetilde{y}$ from $p(\widetilde{y} \mid \theta_1, \theta_2, y)$ where $\widetilde{y}$ is the predictive random variable.

Also, posterior predictive distributions provide a diagnostic check on model specification adequacy. If sample data and posterior predictive draws are substantially different we have evidence of model misspecification.

For conditional simulation, the general procedure for generating posterior predictive draws is

1. draw $\theta_2$ from $p(\theta_2 \mid y)$,
2. draw $\theta_1$ from $p(\theta_1 \mid \theta_2, y)$,
3. draw $\widetilde{y}$ from $p(\widetilde{y} \mid \theta_1, \theta_2, y)$.

## 7.5 Markov chain Monte Carlo (McMC) simulation

Markov chain Monte Carlo (*McMC*) simulations can be employed when the marginal posterior distributions cannot be derived or are extremely cumbersome to derive. *McMC* approaches draw from the set of conditional posterior distributions instead of the marginal posterior distributions. The utility of *McMC* simulation has evolved along with the **R** Foundation for Statistical Computing.

Before discussing standard algorithms (the Gibbs sampler and Metropolis-Hastings) we briefly review some important concepts associated with Markov chains and attempt to develop some intuition regarding their effective usage. The objective is to eventually generate draws from a stationary posterior distribution which we denote $\pi$ but we're unable to directly access. To explore how Markov chains help us access $\pi$, we begin with discrete Markov chains then connect to continuous chains.

## 7.6 Discrete state spaces

Let $S = \left\{\theta^1, \theta^2, \ldots, \theta^d\right\}$ be a discrete state space. A Markov chain is the sequence of random variables, $\{\theta_1, \theta_2, \ldots, \theta_r, \ldots\}$ given $\theta_0$ 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 $\theta_{r+1}$ only depends on the immediate past history, $\theta_r$, 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 $i$.

Now, relate this Markov chain idea to distributions from which random variables are drawn. Say, the initial value, $\theta_0$, is drawn from $\pi_0$. Then, the distribution for $\theta_1$ given $\theta_0 \sim \pi_0$ is

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

In matrix notation, the above is

$$\pi_1^T = \pi_0^T P$$

and after $r$ iterations we have

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

As the number of iterations increases, we expect the effect of the initial distribution, $\pi_0$, dies out so long as the chain does not get trapped.

### 7.6.1 Irreducibility and stationarity

The idea of no absorbing states or states in which the chain gets trapped is called *irreducibility*. This is key to our construction of Markov chains. If $p_{ij} > 0$ (strictly positive) for all $i, j$, then the chain is *irreducible* and there exists a *stationary* distribution, $\pi$, such that

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

and

$$\pi^T P = \pi^T$$

Since the elements are all nonnegative and each row of $P$ sums to one, the maximum eigenvalue of $P^T$ is one and its corresponding eigenvector determines $\pi$. The Perron-Frobenius theorem says for a nonnegative (positive) matrix the largest eigenvalue and its associated eigenvector are real and nonnegative (positive). Further,

$$\iota^T \left(P^T - I\right) = 1 - 1 = 0$$

and the rows of $\left(P^T - I\right)$ sum to the zero row. Hence, $\left(P^T - I\right)$ is singular and one is an eigenvalue. Second, $\left(P^T\right)^r$ is a Markov matrix just as is $P^T$. Therefore, $\pi_r$ cannot grow without bound and the maximum eigenvalue of $P^T$ is one. Putting this together gives, by singular value decomposition, $P^T = S\Lambda S^{-1}$ where $S$ is a matrix of eigenvectors and $\Lambda$ 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 (if $\max \lambda = 1$) and in the direction of its (real and positive) eigenvector (if the remaining $\lambda'$s $< 1$ then $\lambda_i^r$ goes to zero and their corresponding eigenvectors' influence on direction dies out). That is,

$$\left(P^T\right)^r = S_1 \Lambda^r S_1^{-1}$$

where $S_1$ denotes the eigenvector (column vector) corresponding to the unit eigenvalue and $S_1^{-1}$ denotes the corresponding inverse eigenvector (row vector). Since one is the largest eigenvalue of $P^T$, after a large number of iterations $\pi_0^T P^r$ converges to $1 \times \pi = \pi$. Hence, after many iterations the Markov chain produces draws from a stationary distribution if the chain is irreducible.

On the other hand, consider the reducible ($p_{11} = p_{22} = 0$) Markov transition matrix

$$P = P^T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The eigenvalues are 1 and $-1$ with eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The distribution oscillates between the starting distribution vector and its reverse order. For example, suppose $\pi_0^T = \begin{bmatrix} 0.25 & 0.75 \end{bmatrix}$, then $\pi_1^T = \begin{bmatrix} 0.75 & 0.25 \end{bmatrix}$, $\pi_2^T = \begin{bmatrix} 0.25 & 0.75 \end{bmatrix}$, $\pi_3^T = \begin{bmatrix} 0.75 & 0.25 \end{bmatrix}$, and so on. Hence, for this reducible matrix there is no stationary distribution.

One more special case, suppose the Markov transition matrix is irreducible and symmetric, then the stationary distribution is uniform. That is, the eigenvector associated with unit eigenvalue is a vector composed of equal elements. This follows by recognizing $P\iota = 1\iota$ and $P^T = P$ implies $P^T\iota = 1\iota$. Hence, $\iota$ is the eigenvector associated with $\lambda = 1$ and normalization leads to uniform probability assignment.

### *7.6.2 Time reversibility and stationarity*

Another property, *time reversibility*, is sometimes more useful when working with more general state space chains — the chains with which *McMC* methods typically work. 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 $\pi^T P = \pi^T$ where $\pi$ is the stationary distribution for the chain. The reverse transition probability (by Bayesian "updating") is

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

Since the forward chain is Markovian, this simplifies as

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

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 $i$.

The above implies if a chain is reversible with respect to a distribution $\pi$ then $\pi$ is the stationary distribution of the chain. To see this sum both sides of the above relation over $i$

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

In matrix notation, we have

$$\pi^T P = \pi^T$$

$\pi$ is the stationary distribution of the chain.

An important implication of reversibility is the joint distribution $p(\pi, P)$ expressed in matrix form is symmetric.

$$p(\pi, P) = \begin{bmatrix} \pi_1 p_{11} & \pi_1 p_{12} & \cdots & \pi_1 p_{1n} \\ \pi_2 p_{21} & \pi_2 p_{22} & \cdots & \pi_2 p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_n p_{n1} & \pi_n p_{n2} & \cdots & \pi_n p_{nn} \end{bmatrix}$$

We'll revisit the implications of this symmetry when we discuss the Gibbs sampler, specifically, deriving marginals from the set of conditional distributions.

Kolmogorov reversibility criterion

Kolmogorov's theorem provides a mechanism to gauge whether a Markov matrix is reversible without knowing $\pi$. Kolmogorov's criterion says if the product of transition probabilities in a forward loop equals the product of transition probabilities in its reverse loop

$$p_{ij}p_{jk}p_{kl}\cdots p_{ni}=p_{in}\cdots p_{lk}p_{kj}p_{ji}$$

for all loops, then the Markov matrix is reversible. How many loops do we need to check? Every two state matrix is reversible. There is only one three-state loop to check for a three state matrix, three four-state loops for a four state matrix, twelve five-state loops for a five state matrix, and so on. In general, there are $\frac{(n-1)!}{2}$ $n$-state loops for an n state matrix.

Consider a special case where $P$ is defined to be

$$\begin{bmatrix} 1-(n-1)\,e_1 & e_1 & e_1 & \cdots & e_1 \\ e_2 & 1-(n-1)\,e_2 & e_2 & \cdots & e_2 \\ e_3 & e_3 & 1-(n-1)\,e_3 & & e_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_n & e_n & e_n & \cdots & 1-(n-1)\,e_n \end{bmatrix}$$

As every $n$-element loop involves one non-diagonal element from each row, Kolmogorov's criterion is satisfied and such Markov transition matrices are reversible. A similar (special case) reversible transition matrix involves the same non-diagonal elements in each column.

$$\begin{bmatrix} \begin{array}{c}1-e_2-e_3\\ \cdots-e_n\end{array} & e_2 & e_3 & \cdots & e_n \\ e_1 & \begin{array}{c}1-e_1-e_3\\ \cdots-e_n\end{array} & e_3 & \cdots & e_n \\ e_1 & e_2 & \begin{array}{c}1-e_1-e_2\\ \cdots-e_n\end{array} & & e_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_1 & e_2 & e_3 & \cdots & \begin{array}{c}1-e_1-e_2\\ \cdots-e_{n-1}\end{array} \end{bmatrix}$$

## 7.7 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, $\theta$. That is, the kernel of the chain, $K(\theta, A)$, is the probability of set $A$ given the chain is at $\theta$ where

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

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

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 $\pi$ 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 $K^n(\theta, A) > 0$. With continuous state spaces, irreducibility and time reversibility produce a stationary distribution of the chain as with discrete state spaces.

Next, we briefly discussion application of these Markov chain concepts to two popular *McMC* strategies: the Gibbs sampler, and Metropolis-Hastings (*MH*) algorithm. The Gibbs sampler is a special case of *MH* and somewhat simpler so we review it first.

## 7.8 Gibbs sampler

Suppose we cannot derive $p(\theta \mid Y)$ 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.[4]

[4] The Gibbs sampler is a Markov chain Monte Carlo simulation approach developed by Geman and Geman [1984] to address the Gibbs distribution (also Boltzman distribution) in Markov networks. The Gibbs measure is the probability of system $X$ being in state $x$

$$P(X = x) = \frac{\exp(-\beta E[x])}{Z(\beta)}$$

with physical interpretation $E[x]$ is energy of $x$, $\beta$ is inverse temperature, and $Z(\beta)$ is the normalizing partition function. We see the Gibbs measure is the foundation of maximum entropy probability assignment. The Hammersley-Clifford theorem (a foundational idea of McMC analyses along with Besag [1974]) says the Gibbs measure can represent the distribution of any Markovian process made up of only positive probabilities.

For the full set of conditional posterior distributions

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

draws are made for $\theta_1$ conditional on starting values for parameters other than $\theta_1$, that is $\theta_{-1}$. Then, $\theta_2$ is drawn conditional on the $\theta_1$ draw and the starting values for the remaining $\theta$. Next, $\theta_3$ is drawn conditional on the draws for $\theta_1$ and $\theta_2$ and the starting values for the remaining $\theta$. This continues until all $\theta$ 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 $\theta_2$,
2. draw $\theta_1$ from $p\left(\theta_1|\theta_2, y\right)$ utilizing the starting value for $\theta_2$,
3. draw $\theta_2$ from $p(\theta_2|\theta_1, y)$ utilizing the previous draw for $\theta_1$,
4. repeat until a converged sample based on the marginal posteriors is obtained.

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 (see Gelfand and Smith [1990]). (Note, in general, $p\left(\theta_1, \theta_2 \mid Y\right) \neq p\left(\theta_1 \mid \theta_2, Y\right) p\left(\theta_2 \mid \theta_1, Y\right)$.) 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.

### *7.8.1 Marginals from set of conditionals*

The Gibbs sampler derives the desired (approximate, in finite samples) marginal posterior distributions from the set of conditional posterior distributions. While deriving marginals from a joint distribution is simply the sum rule, the idea of deriving marginals from conditionals is a bit more involved. It is feasible to derive marginals from conditionals provided the joint distribution exists. We illustrate its feasibility for the two variable/parameter case. We'll focus on $X$ but the analogous ideas apply to $Y$. From the sum rule we have

$$f_X\left(x\right) = \int f_{X,Y}\left(x, y\right) dy$$

while the product rule gives

$$f_X(x) = \int f_{X|Y}(x \mid y)\, f_Y(y)\, dy$$

Now, apply the product rule again where we think of $X_{k-1}$ rather than draw $k$. This yields

$$f_X(x) = \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t)\, f_X(t)\, dtdy$$

Next, rearrangement leads to

$$f_X(x) = \int \left[\int f_{X|Y}(x \mid y)\, f_{Y|X}(y \mid t)\, dy\right] f_X(t)\, dt$$

As $y$ is integrated out, the term in brackets is a function of $x$ and $t$, $h(x,t) = \left[\int f_{X|Y}(x \mid y)\, f_{Y|X}(y \mid t)\, dy\right]$. Hence,

$$f_X(x) = \int h(x,t)\, f_X(t)\, dt$$

which is a fixed point integral equation for which $f_X(x) = f_X(t)$ for $x = t$ is a unique solution. Below we illustrate this result for a simple, but illuminating discrete case.

### 7.8.2 Example

Suppose the unknown joint and marginal distributions are as follows.

| $f_{X,Y}(x,y)$ | $y=4$ | $y=5$ | $y=6$ | $f_X(x)$ |
|---|---|---|---|---|
| $x=1$ | 0.1 | 0.05 | 0.15 | 0.3 |
| $x=2$ | 0.15 | 0.2 | 0.05 | 0.4 |
| $x=3$ | 0.05 | 0.1 | 0.15 | 0.3 |
| $f_Y(y)$ | 0.3 | 0.35 | 0.35 | |

The known conditional distributions are

| $f_{X\|Y}(x \mid y)$ | $y=4$ | $y=5$ | $y=6$ |
|---|---|---|---|
| $x=1$ | $\frac{1}{3}$ | $\frac{1}{7}$ | $\frac{3}{7}$ |
| $x=2$ | $\frac{1}{2}$ | $\frac{4}{7}$ | $\frac{1}{7}$ |
| $x=3$ | $\frac{1}{6}$ | $\frac{2}{7}$ | $\frac{3}{7}$ |

and

| $f_{Y\|X}(y \mid x)$ | $y=4$ | $y=5$ | $y=6$ |
|---|---|---|---|
| $x=1$ | $\frac{1}{3}$ | $\frac{1}{6}$ | $\frac{1}{2}$ |
| $x=2$ | $\frac{3}{8}$ | $\frac{1}{2}$ | $\frac{1}{8}$ |
| $x=3$ | $\frac{1}{6}$ | $\frac{1}{3}$ | $\frac{1}{2}$ |

The product of the conditionals summed over $y$ yields

| $h(x,t)$ | $t=1$ | $t=2$ | $t=3$ |
|---|---|---|---|
| $x=1$ | 0.3492 | 0.2500 | 0.3175 |
| $x=2$ | 0.3333 | 0.4911 | 0.3452 |
| $x=3$ | 0.3175 | 0.2589 | 0.3373 |

where $h(x,t) = \sum_y f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)$. As written, $h(x,t)$ is the transpose of the Markov transition matrix, $P^T = h(x,t)$ where each column sums to one. It has maximum eigenvalue equal to one and associated eigenvector proportional to $\pi^T = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}^T$. Hence, from the set of conditional posteriors we have recovered the transition matrix and the stationary distribution $\pi$.

To complete the discussion started above and verify the immediately preceding claim, we solve for the marginal distribution of $X$, $\pi \equiv f_X(x)$, by equating $f_X(x) = \sum_t h(x,t) f_X(t)$. Marginalization of $X$ involves $f_X(t)$ but it is unknown and cannot be directly retrieved from $h(x,t)$ since it is not a proper density (or mass) function (it sums to 3 and $h(t) = \sum_x h(x,t) = 1$ for all $t = 1,2,3$ — recall these are transition probabilities). However, we can solve $f_X(x) = \sum_t h(x,t) f_X(t)$ such that $f_X(x) = f_X(t)$ for $x = t$ and $\sum_x f_X(x) = 1$.

| $h(x,t) f_X(t)$ | $t=1$ | $t=2$ | $t=3$ |
|---|---|---|---|
| $x=1$ | 0.1048 | 0.1000 | 0.0952 |
| $x=2$ | 0.1000 | 0.1964 | 0.1036 |
| $x=3$ | 0.0952 | 0.1036 | 0.1012 |

Notice, this matrix exhibits the time reversible symmetry we alluded to earlier in the discussion of time reversibility where the (unique) fixed point solution from $f_X(x) = \sum_t h(x,t) f_X(t)$ is

$$f_X(x) = f_X(t) = \begin{array}{ll} 0.3, & x=1 \\ 0.4, & x=2 \\ 0.3, & x=3 \end{array}$$

the desired marginal distribution.[5] Therefore, when we take a sufficiently large number of draws from the conditionals we reach (approximately, in finite samples) $p(\pi, P)$ and numerous draws from this distribution (covering $X_{k-1} = t$ for almost all $t$) reveals the marginal distribution for $X$.

[5] Applying the analogous ideas to derive the marginal distribution for $Y$, we have

| $h(y,t)$ | $t=4$ | $t=5$ | $t=6$ |
|---|---|---|---|
| $y=4$ | 0.3264 | 0.3095 | 0.2679 |
| $y=5$ | 0.3611 | 0.4048 | 0.2857 |
| $y=6$ | 0.3125 | 0.2857 | 0.4464 |

### *7.8.3 more general approach*

For more than two blocks of variables/parameters for which we desire their marginal posterior distributions but only know the set of conditional distributions we follow Besag [1974]. Suppose we have $n$ blocks of variables: $X = X_1, \ldots, X_n$. and we have two sets of realizations $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ for these variables, say, draw $k$ and $k-1$ as before. The joint distribution can be written

$$p(x_1, \ldots, x_n) = p(x_1 \mid x_2, \ldots, x_n)\, p(x_2, \ldots, x_n)$$

but we know the first term on the right hand side but not the second term. Next, add $k-1$ draws for $X_1$ in the form $\frac{p(x_2,\ldots,x_n,y_1)}{p(y_1 \mid x_2,\ldots,x_n)} = p(x_2, \ldots, x_n)$ and write

$$p(x_1, \ldots, x_n) = \frac{p(x_1 \mid x_2, \ldots, x_n)\, p(x_2, \ldots, x_n, y_1)}{p(y_1 \mid x_2, \ldots, x_n)}$$

Again, apply the above procedure for the problematic term in the numerator by adding $k-1$ draws for $X_2$ by writing

$$p(x_2, \ldots, x_n, y_1) = \frac{p(x_2 \mid x_3, \ldots, x_n, y_1)\, p(x_3, \ldots, x_n, y_1, y_2)}{p(y_2 \mid x_3, \ldots, x_n, y_1)}$$

Substitution and continuing the procedure yields

$$\begin{aligned} p(x_1, \ldots, x_n) &= \frac{p(x_1 \mid x_2, \ldots, x_n)}{p(y_1 \mid x_2, \ldots, x_n)} \frac{p(x_2 \mid x_3, \ldots, x_n, y_1)}{p(y_2 \mid x_3, \ldots, x_n, y_1)} \\ &\quad \times p(x_3, \ldots, x_n, y_1, y_2) \\ &= \frac{p(x_1 \mid x_2, \ldots, x_n)}{p(y_1 \mid x_2, \ldots, x_n)} \frac{p(x_2 \mid x_3, \ldots, x_n, y_1)}{p(y_2 \mid x_3, \ldots, x_n, y_1)} \\ &\quad \cdots \frac{p(x_n \mid y_1, \ldots, y_{n-1})}{p(y_n \mid y_1, \ldots, y_{n-1})} p(y_1, \ldots, y_n) \end{aligned}$$

Now, we have the ratio of the joint distribution for two different sets of realizations equal to the product of ratios of the known conditional distri-

for the (transpose of the) transition matrix with characteristic vector, $\pi^T = \begin{bmatrix} 0.3 & 0.35 & 0.35 \end{bmatrix}$,

| $h(y,t)\, f_Y(t)$ | $t=4$ | $t=5$ | $t=6$ |
|---|---|---|---|
| $y=4$ | 0.09792 | 0.10833 | 0.09375 |
| $y=5$ | 0.10833 | 0.14172 | 0.10000 |
| $y=6$ | 0.09375 | 0.10000 | 0.15625 |

and

$$f_Y(y) = f_Y(t) = \begin{array}{ll} 0.3, & y=4 \\ 0.35, & y=5 \\ 0.35, & y=6 \end{array}$$

butions.

$$\frac{p(x_1,\ldots,x_n)}{p(y_1,\ldots,y_n)} = \frac{p(x_1 \mid x_2,\ldots,x_n)}{p(y_1 \mid x_2,\ldots,x_n)} \frac{p(x_2 \mid x_3,\ldots,x_n,y_1)}{p(y_2 \mid x_3,\ldots,x_n,y_1)} \cdots \frac{p(x_n \mid y_1,\ldots,y_{n-1})}{p(y_n \mid y_1,\ldots,y_{n-1})}$$

How does this aid our quest for the marginal distribution? If we think about this the ratios of the joint contain a sufficient amount of information for deducing the joint distribution. We'll illustrate this by way of a numerical example.

### 7.8.4 *example*

Suppose we have 3 variables with the following unknown joint and marginal distributions.

| $p(x_1,x_2,x_3)$ | $x_2,x_3$ | $x_2,x_3$ | $x_2,x_3$ | $x_2,x_3$ | $p(x_1)$ |
|---|---|---|---|---|---|
| | $=1,1$ | $=1,2$ | $=2,1$ | $=2,2$ | |
| $x_1=1$ | 0.10 | 0.05 | 0.15 | 0.05 | 0.35 |
| $x_1=2$ | 0.10 | 0.05 | 0.05 | 0.10 | 0.30 |
| $x_1=3$ | 0.05 | 0.10 | 0.05 | 0.15 | 0.35 |

| | $p(x_2)$ | $p(x_3)$ |
|---|---|---|
| $x_j=1$ | 0.45 | 0.50 |
| $x_j=2$ | 0.55 | 0.50 |

We know the conditional distributions

| $p(x_1 \mid x_2,x_3)$ | $x_2,x_3$ | $x_2,x_3$ | $x_2,x_3$ | $x_2,x_3$ |
|---|---|---|---|---|
| | $=1,1$ | $=1,2$ | $=2,1$ | $=2,2$ |
| $x_1=1$ | 0.40 | 0.25 | 0.60 | $\frac{1}{6}$ |
| $x_1=2$ | 0.40 | 0.25 | 0.20 | $\frac{1}{3}$ |
| $x_1=3$ | 0.20 | 0.50 | 0.20 | $\frac{1}{2}$ |

| $p(x_2 \mid x_1,x_3)$ | $x_1,x_3$ | $x_1,x_3$ | $x_1,x_3$ | $x_1,x_3$ | $x_1,x_3$ | $x_1,x_3$ |
|---|---|---|---|---|---|---|
| | $=1,1$ | $=1,2$ | $=2,1$ | $=2,2$ | $=3,1$ | $=3,2$ |
| $x_2=1$ | 0.40 | 0.50 | $\frac{2}{3}$ | $\frac{1}{3}$ | 0.50 | 0.40 |
| $x_2=2$ | 0.60 | 0.50 | $\frac{1}{3}$ | $\frac{2}{3}$ | 0.50 | 0.60 |

| $p(x_3 \mid x_1,x_2)$ | $x_1,x_2$ | $x_1,x_2$ | $x_1,x_2$ | $x_1,x_2$ | $x_1,x_2$ | $x_1,x_2$ |
|---|---|---|---|---|---|---|
| | $=1,1$ | $=1,2$ | $=2,1$ | $=2,2$ | $=3,1$ | $=3,2$ |
| $x_3=1$ | $\frac{2}{3}$ | 0.75 | $\frac{2}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | 0.25 |
| $x_3=2$ | $\frac{1}{3}$ | 0.25 | $\frac{1}{3}$ | $\frac{2}{3}$ | $\frac{2}{3}$ | 0.75 |

From the conditional distributions we derive the ratio of the joint distribution where $x_j$ denotes the realization from draw $k$ and $y_j$ denotes the realization from draw $k-1$. For example,

$$\begin{aligned}
& \frac{p\left(x_1=1,x_2=1,x_3=1\right)}{p\left(y_1=1,y_2=1,y_3=2\right)} \\
= & \frac{p\left(x_1=1 \mid x_2=1,x_3=1\right)}{p\left(y_1=1 \mid x_2=1,x_3=1\right)} \frac{p\left(x_2=1 \mid x_3=1,y_1=1\right)}{p\left(y_2=1 \mid x_3=1,y_1=1\right)} \\
& \times \frac{p\left(x_3=1 \mid y_1=1,y_2=1\right)}{p\left(y_3=2 \mid y_1=1,y_2=1\right)} \\
= & \frac{0.40}{0.40} \frac{0.40}{0.40} \frac{\frac{2}{3}}{\frac{1}{3}} = 2
\end{aligned}$$

It is sufficient, for example, to derive the complete set of ratios involving $\frac{p(x_1=1,x_2=1,x_3=1)}{p(y_1,y_2,y_3)}$ for all values of $y_1, y_2$, and $y_3$. Then, form the linear equation based on the relative likelihoods and the sum of the likelihoods equal to one

$$Ap = b$$

where $p$ is a vector completely describing the joint distribution for $X$,

$$A = \begin{array}{cccccccccccc}
1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -\frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{2}{3} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

and

$$b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Since $A$ is full rank the solution is the unique joint distribution for $X$.

$$p = \begin{bmatrix} p\left(x_1 = 1, x_2 = 1, x_3 = 1\right) \\ p\left(x_1 = 1, x_2 = 1, x_3 = 2\right) \\ p\left(x_1 = 1, x_2 = 2, x_3 = 1\right) \\ p\left(x_1 = 1, x_2 = 2, x_3 = 2\right) \\ p\left(x_1 = 2, x_2 = 1, x_3 = 1\right) \\ p\left(x_1 = 2, x_2 = 1, x_3 = 2\right) \\ p\left(x_1 = 2, x_2 = 2, x_3 = 1\right) \\ p\left(x_1 = 2, x_2 = 2, x_3 = 2\right) \\ p\left(x_1 = 3, x_2 = 1, x_3 = 1\right) \\ p\left(x_1 = 3, x_2 = 1, x_3 = 2\right) \\ p\left(x_1 = 3, x_2 = 2, x_3 = 1\right) \\ p\left(x_1 = 3, x_2 = 2, x_3 = 2\right) \end{bmatrix} = \begin{bmatrix} 0.10 \\ 0.05 \\ 0.15 \\ 0.05 \\ 0.10 \\ 0.05 \\ 0.05 \\ 0.10 \\ 0.05 \\ 0.10 \\ 0.05 \\ 0.15 \end{bmatrix}$$

Gaussian elimination and back substitution applied to $Ap = b$ reveals a deeper result regarding the full matrix, $B$, of joint likelihood ratios, $\frac{p(x_1,x_2,x_3)}{p(y_1,y_2,y_3)}$.

$$B = \begin{bmatrix} \frac{p(1,1,1)}{p(1,1,1)} & \frac{p(1,1,1)}{p(1,1,2)} & \frac{p(1,1,1)}{p(1,2,1)} & \cdots & \frac{p(1,1,1)}{p(3,2,2)} \\ \frac{p(1,1,2)}{p(1,1,1)} & \frac{p(1,1,2)}{p(1,1,2)} & \frac{p(1,1,2)}{p(1,2,1)} & \cdots & \frac{p(1,1,2)}{p(3,2,2)} \\ \frac{p(1,1,3)}{p(1,1,1)} & \frac{p(1,1,3)}{p(1,1,2)} & \frac{p(1,1,3)}{p(1,2,1)} & \cdots & \frac{p(1,1,3)}{p(3,2,2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{p(3,2,2)}{p(1,1,1)} & \frac{p(3,2,2)}{p(1,1,2)} & \frac{p(3,2,2)}{p(1,2,1)} & \cdots & \frac{p(3,2,2)}{p(3,2,2)} \end{bmatrix}$$

Each row of $B$ identified by the numerator of the ratio, $p\left(l, m, n\right)$, can be utilized to create a matrix $A$ from which row operations reveal

$$p\left(i, j, k\right) = \frac{\frac{p(i,j,k)}{p(l,m,n)}}{Z\left(l, m, n\right)}$$

where $Z(l, m, n) = \sum_{r,s,t} \frac{p(r,s,t)}{p(l,m,n)}$. In other words, each column of $B$ (identified by the denominator of the ratio, $p(l, m, n)$) rescaled (normalized) by the column sum identifies the joint distribution, $p(x_1, x_2, x_3)$.

These concepts reinforce the notion that generating simulated draws from the relevant marginal posterior distribution based on the set of conditional posterior distributions entails numerous (Gibbs) samples.

## 7.9 Metropolis-Hastings algorithm

If neither some conditional posterior, $p(\theta_j \mid Y, \theta_{-j})$, or its marginal posterior, $p(\theta \mid Y)$, is recognizable, then we may be able to employ the Metropolis-Hastings (*MH*) algorithm. The Gibbs sampler is a special case of the *MH* algorithm. The random walk Metropolis algorithm is most common and outlined next.

We wish to draw from $p(\theta \mid \cdot)$ but we only know $p(\theta \mid \cdot)$ up to constant of proportionality, $p(\theta \mid \cdot) = cf(\theta \mid \cdot)$ where $c$ is unknown. The random walk Metropolis algorithm for one parameter is as follows.

1. Let $\theta^{(k-1)}$ be a draw from $p(\theta \mid \cdot)$.
2. Draw $\theta^*$ from $N\left(\theta^{(k-1)}, s^2\right)$ where $s^2$ is fixed.
3. Let $\alpha = min\left\{1, \frac{p(\theta^* \mid \cdot)}{p\left(\theta^{(k-1)} \mid \cdot\right)} = \frac{cf(\theta^* \mid \cdot)}{cf\left(\theta^{(k-1)} \mid \cdot\right)}\right\}$.
4. Draw $z^*$ from $U(0, 1)$.
5. If $z^* < \alpha$ then $\theta^{(k)} = \theta^*$, otherwise $\theta^{(k)} = \theta^{(k-1)}$. In other words, with probability $\alpha$ set $\theta^{(k)} = \theta^*$, and otherwise set $\theta^{(k)} = \theta^{(k-1)}$.[6]

These draws converge to random draws from the marginal posterior distribution after a burn-in interval if properly tuned.

Tuning the Metropolis algorithm involves selecting $s^2$ (jump size) so that the parameter space is explored appropriately. Usually, smaller jump size results in more accepts and larger jump size results in fewer accepts. If $s^2$ is too small, the Markov chain will not converge quickly, has more serial correlation in the draws, and may get stuck at a local mode (multi-modality can be a problem). If $s^2$ is too large, the Markov chain will move around too much and not be able to thoroughly explore areas of high posterior probability. Of course, we desire concentrated samples from the posterior distribution. A commonly-employed rule of thumb is to target an acceptance rate for $\theta^*$ around 30% ($20 - 80\%$ is usually considered "reasonable").[7]

[6] A modification of the *RW* Metropolis algorithm sets $\theta^{(k)} = \theta^*$ with $log(\alpha)$ probability where $\alpha = min\{0, log[f(\theta^* \mid \cdot)] - log[f(\theta^{(k-1)} \mid \cdot)]\}$.

[7] Gelman, et al [2004] report the optimal acceptance rate is 0.44 when the number of parameters $K = 1$ and drops toward 0.23 as $K$ increases.

The above procedure describes the algorithm for a single parameter or vector of parameters. A general $K$ parameter Metropolis-Hastings algorithm works similarly (see Train [2002], p. 305).

1. Start with a value $\beta_n^0$.

2. Draw $K$ independent values from a standard normal density, and stack the draws into a vector labeled $\eta^1$.

3. Create a trial value of $\beta_n^1 = \beta_n^0 + \sigma\Gamma\eta^1$ where $\sigma$ is the researcher-chosen jump size parameter, $\Gamma$ is the Cholesky factor of $W$ such that $\Gamma\Gamma^T = W$. Note the proposal distribution is specified to be normal with zero mean and variance $\sigma^2 W$.

4. Draw a standard uniform variable $\mu_1$.

5. Calculate the ratio $F = \frac{L\left(\beta_n^1|y_n\right)\phi\left(\beta_n^1\right)}{L(\beta_n^0|y_n)\phi(\beta_n^0)}$ where $L\left(\beta_n^1 \mid y_n\right)$ is the likelihood for the proposal and $\phi\left(\beta_n^1\right)$ is the prior for the proposal and $L\left(\beta_n^0 \mid y_n\right)$ and $\phi\left(\beta_n^0\right)$ are the analogs for the initial or previous draw.

6. If $\mu_1 \leq F$, accept $\beta_n^1$; if $\mu_1 > F$ , reject $\beta_n^1$ and let $\beta_n^1 = \beta_n^0$.

7. Repeat the process many times, adjusting the tuning parameters if necessary. For sufficiently large $t$, $\beta_n^t$ is a draw from the marginal posterior.

### *7.9.1 Metropolis-Hastings algorithm and reversibility*

Now, we explore how the *MH* algorithm exploits reversibility to assure that a convergent distribution $\pi^*$ can be found. Since we don't have the full set of conditional posterior distributions we cannot directly derive the transition kernel, rather we choose a candidate generator density (for the next potential draw $j$) $q_{ij}$ and adapt it, if necessary, to satisfy reversibility. If $\pi_i q_{ij} = \pi_j q_{ji}$, that is reversibility is satisfied, then our search is complete. However, this is unlikely. Rather, suppose $\pi_i q_{ij} > \pi_j q_{ji}$. This indicates instability as the process moves from $i$ to $j$ too often and too infrequently from $j$ to $i$. To adjust for this we introduce a probability of move parameter, $\alpha_{ij} < 1$, to reduce the frequency of moves from $i$ to $j$. Transitions from $i$ to $j$ are made according to $p_{ij}^{MH} = q_{ij}\alpha_{ij}$. If no transition occurs the generator returns the previous draw $i$. $\alpha_{ij}$ is determined to satisfy the reversibility condition.

$$\pi_i q_{ij} \alpha_{ij} = \pi_j q_{ji} \alpha_{ji}$$

Since $\pi_i q_{ij} > \pi_j q_{ji}$ tells us the process too infrequently moves from $j$ to $i$, we make $\alpha_{ji} = 1$, as large as possible (one, since it's a probability). This leaves

$$\begin{aligned} \pi_i q_{ij} \alpha_{ij} &= \pi_j q_{ji} \\ \alpha_{ij} &= \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \end{aligned}$$

As $\pi$ is only proportional to the distribution of interest ($f = c\pi$, for some unknown constant $c$), define the probability of move as

$$\alpha_{ij} = \begin{array}{cc} \min\left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right) & \pi_i q_{ij} > 0 \\ 1 & otherwise \end{array}$$

Since $\alpha_{ij}$ is a ratio, the normalizing constant $c$ cancels. If the generator density $q$ is symmetric, $q_{ij} = q_{ji}$, and $\alpha_{ij}$ simplifies to

$$\alpha_{ij} = \begin{array}{cc} \min\left(\frac{\pi_j}{\pi_i}, 1\right) & \pi_i q_{ij} > 0 \\ 1 & otherwise \end{array}$$

so that if $\pi_j > \pi_i$ the process moves to $j$, otherwise it moves with probability $\frac{\pi_j}{\pi_i}$. This standard version of the *MH* algorithm accepts an $\alpha$ fraction of moves from $i$ to $j$ and effectively all moves from $j$ to $i$. (i.e., rejecting a move is as if a move was made from $j$ to $i$).

On the other hand, suppose $\pi_i q_{ij} < \pi_j q_{ji}$, then time reversibility is ensured if we reverse the above process. That is,

$$\pi_i q_{ij} \alpha_{ij} = \pi_j q_{ji} \alpha_{ji}$$

allows us to set the probability of a forward move $\alpha_{ij} = 1$ and control the probability of a reverse move by $\alpha_{ji} = \min\left(\frac{\pi_i q_{ij}}{\pi_j q_{ji}}, 1\right)$. Therefore, collectively we have

$$\pi_i p_{ij} = \pi_i q_{ij} \alpha_{ij} = \pi_i q_{ij} \min\left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right) = \min\left(\pi_j q_{ji}, \pi_i q_{ij}\right)$$

$$\pi_j p_{ji} = \pi_j q_{ji} \alpha_{ji} = \pi_j q_{ji} \min\left(\frac{\pi_i q_{ij}}{\pi_j q_{ji}}, 1\right) = \min\left(\pi_j q_{ji}, \pi_i q_{ij}\right)$$

In other words, reversibility is satisfied, $\pi_i p_{ij} = \pi_j p_{ji}$, and administering either the forward or reverse chain achieves the same result.

The generator density $q$ determines the *MH* algorithm with transition controlled via $\alpha_{ij}$. The algorithm effectively defines a transition kernel $K$ whose $n$th iterate converges to the target distribution $\pi$ for large $n$. The *MH* algorithm ensures reversibility by creating a Markov chain which moves with probability $\alpha$ and repeats the last draw with probability $1 - \alpha$. If the reversibility property is satisfied, then $\pi$ is the invariant distribution for $K$ (for additional details, see Tierney [1994], Chib and Greenberg [1995], and below).

### MH and kernel convergence

Following the continuous state spaces notation and setup we have a transition kernel, as a function of a generator density $p(\theta, \phi)$, defined as

$$K(\theta, \phi) = p(\theta, \phi)\, d\phi + r(\theta)\, \delta_\theta(\phi)$$

where $p(\theta,\theta) = 0$, $\delta_\theta(\phi) = 1$ if $\theta \in d\phi$ and 0 otherwise, and $r(\theta) = 1 - \int p(\theta,\phi)\,d\phi$, the probability the chain remains at $\theta$. The $n$th iterate of the transition kernel is given by

$$K^{(n)}(\theta, A) = \int K^{(n-1)}(\theta, d\phi)\,K(\phi, A)$$

where $K^{(1)}(\theta, d\phi) = K(\theta, d\phi)$. Invariance in distribution implies

$$\pi^*(d\phi) = \int K(\theta, d\phi)\,\pi(\theta)\,d\theta$$

For *McMC* methods, the invariant distribution $\pi(\cdot)$ is known (up to a constant) but the kernel $K(\cdot,\cdot)$ is unknown. The task is to define an algorithm that identifies a transition kernel. The above discussion suggests the kernel is primarily determined by the generator density $p(\cdot,\cdot)$. Next, we show that if the generator is reversible, then the known target distribution is the invariant distribution associated with the kernel. In other words, we can identify an algorithm (*MH*) for determining the transition kernel.

If $p(\cdot,\cdot)$ satisfies reversibility $p(\theta,\phi)\,\pi(\theta) = p(\phi,\theta)\,\pi(\phi)$, then $\pi(\cdot)$ is the invariant distribution for $K(\cdot,\cdot)$.

$$\begin{aligned}
\int K(\theta, A)\,\pi(\theta)\,d\theta &= \int \left[\int_A p(\theta,\phi)\,d\phi\right] \pi(\theta)\,d\theta \\
&\quad + \int r(\theta)\,\delta_\theta(A)\,\pi(\theta)\,d\theta \\
&= \int_A \left[\int p(\theta,\phi)\,\pi(\theta)\,d\theta\right] d\phi \\
&\quad + \int_A r(\theta)\,\pi(\theta)\,d\theta \\
&= \int_A \left[\int p(\phi,\theta)\,\pi(\phi)\,d\theta\right] d\phi \\
&\quad + \int_A r(\theta)\,\pi(\theta)\,d\theta \\
&= \int_A (1 - r(\phi))\,\pi(\phi)\,d\phi + \int_A r(\theta)\,\pi(\theta)\,d\theta \\
&= \int_A \pi(\phi)\,d\phi
\end{aligned}$$

The left hand side gives the probability of transitioning from $\theta$ to $\phi$, where $\theta$ is generated by $\pi(\cdot)$, while the right hand side gives the probability of moving from $\phi$ to $\theta$,. again where $\phi$ is generated by $\pi(\cdot)$. Reversibility assures equality of the two sides. In summary, the *MH* algorithm utilizes a generator density to define the transition kernel associated with the known invariant distribution where (kernel) convergence occurs after a large number of iterations.

### *7.9.2 Gibbs sampler as a special case of the MH algorithm*

For the Gibbs sampler, draws are generated directly from the conditional posteriors and the *MH* acceptance probability is $\alpha_{ij} = 1$ for all $i$ and $j$. In other words, the (conditional) transition kernels are defined by the conditional posterior distributions — each identified in blocks. The key is to recognize that "block-at-a-time" sampling from conditional distributions converges to the invariant joint distribution $\pi^*$. (Recall, if the joint or marginal distributions were identifiable, we could use them directly for posterior simulation.) This is a tremendous practical advantage as it allows us to take draws from each block in succession rather than having to run each block to convergence for every conditioning value (see Hastings [1970], Chib and Greenberg [1995], and below for details).

Block sampling and conditional kernel convergence

We continue with the notation and setup utilized in the kernel convergence discussion above. Define two blocks as $\theta = (\theta_1, \theta_2)$. Suppose there exists a conditional transition kernel $K_1(\theta_1, d\phi_1 \mid \theta_2)$ such that $\pi^*_{1|2}(\cdot \mid \theta_2)$ is its invariant conditional distribution.

$$\pi^*_{1|2}(d\phi_1 \mid \theta_2) = \int K_1(\theta_1, d\phi_1 \mid \theta_2)\, \pi_{1|2}(\theta_1 \mid \theta_2)\, d\theta_1$$

The analogous conditional kernel exists for the other block

$$\pi^*_{2|1}(d\phi_2 \mid \theta_1) = \int K_2(\theta_2, d\phi_2 \mid \theta_1)\, \pi_{2|1}(\theta_2 \mid \theta_1)\, d\theta_2$$

As alluded to above, the key result is the product of conditional kernels has the joint distribution $\pi(\cdot,\cdot)$ as its invariant distribution. The result is demonstrated below. Suppose $K_1(\cdot,\cdot \mid \theta_2)$ produces $\phi_1$ given $\theta_1$ and $\theta_2$, and $K_2(\cdot,\cdot \mid \phi_1)$ generates $\phi_2$ given $\theta_2$ and $\phi_1$.

$$\begin{aligned}
& \int\int K_1(\theta_1, d\phi_1 \mid \theta_2)\, K_2(\theta_2, d\phi_2 \mid \phi_1)\, \pi(\theta_1, \theta_2)\, d\theta_1 d\theta_2 \\
= & \int K_2(\theta_2, d\phi_2 \mid \phi_1) \left[\int K_1(\theta_1, d\phi_1 \mid \theta_2)\, \pi_{1|2}(\theta_1 \mid \theta_2)\, d\theta_1\right] \\
& \times \pi_2(\theta_2)\, d\theta_2 \\
= & \int K_2(\theta_2, d\phi_2 \mid \phi_1)\, \pi^*_{1|2}(d\phi_1 \mid \theta_2)\, \pi_2(\theta_2)\, d\theta_2 \\
= & \int K_2(\theta_2, d\phi_2 \mid \phi_1)\, \frac{\pi_{2|1}(\theta_2 \mid \phi_1)\, \pi^*_1(d\phi_1)}{\pi_2(\theta_2)}\, \pi_2(\theta_2)\, d\theta_2 \\
= & \pi^*_1(d\phi_1) \int K_2(\theta_2, d\phi_2 \mid \phi_1)\, \pi_{2|1}(\theta_2 \mid \phi_1)\, d\theta_2 \\
= & \pi^*_1(d\phi_1)\, \pi^*_{2|1}(d\phi_2 \mid \phi_1) \\
= & \pi^*(d\phi_1, d\phi_2)
\end{aligned}$$

This result sets the stage for a variety of block-at-a-time *McMC* sampling schemes.

## 7.10 Missing data augmentation

One of the many strengths of *McMC* approaches is their flexibility for dealing with missing data. Missing data is a common characteristic plaguing limited dependent variable models like discrete choice and selection. As a prime example, we next discuss Albert and Chib's *McMC* data augmentation approach to discrete choice modeling. Later we'll explore *McMC* data augmentation of selection models.

### *7.10.1 Albert and Chib's Gibbs sampler Bayes' probit*

The challenge with discrete choice models (like probit) is that latent utility is unobservable, rather the analyst observes only discrete (usually binary) choices.[8] Albert & Chib [1993] employ Bayesian data augmentation to "supply" the latent variable. Hence, parameters of a probit model are estimated via normal Bayesian regression (see earlier discussion in this chapter). Consider the latent utility model

$$U_D = W\theta - V$$

where binary choice, $D$, is observed.

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

The conditional posterior distribution for $\theta$ 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 $\theta$ and $Q = \left(W_0^T W_0\right)^{-1}$ is the prior for the covariance. The conditional posterior distribution for the latent variables are

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

[8] See *Accounting and causal effects: econometric challenges*, chapter 5 for a discussion of discrete choice models.

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

where $TN\left(\cdot\right)$ refers to random draws from a truncated normal (truncated below for the first and truncated above for the second). Iterative draws for $(U_D|D,W,\theta)$ and $(\theta|D,W,U_D)$ form the Gibbs sampler. Interval estimates of $\theta$ are supplied by post-convergence draws of $(\theta|D,W,U_D)$. For simulated normal draws of the unobservable portion of utility, $V$, this Bayes' augmented data probit produces remarkably similar inferences to *MLE*.[9]

### 7.10.2 Probit example

We compare *ML* (maximum likelihood) estimates[10] with Gibbs sampler *McMC* data augmentation probit estimates for a simple discrete choice problem. In particular, we return to the choice (or selection) equation referred to in the illustration of the control function strategy for identifying treatment effects of the projection chapter. The variables (choice and instruments) are

| $D$ | $Z_1$ | $Z_2$ | $Z_3$ | $Z_4$ | $Z_5$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 3 | 1 | 1 | −2 |
| 1 | −1 | 0 | 0 | 0 | 1 |
| 0 | −1 | 0 | 0 | 2 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 2 | 0 | 0 | 0 |

The above data are a representative sample. To mitigate any small sample bias, we repeat this sample 20 times ($n=120$).[11]

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

$$E\left[U_D\mid Z\right]=-\underset{(0.2095)}{0.6091}Z_1+\underset{(0.1454)}{0.4950}Z_2-\underset{(0.2618)}{0.1525}Z_3-\underset{(0.1922)}{0.7233}Z_4+\underset{(0.1817)}{0.2283}Z_5$$

The model has only modest explanatory power (pseudo-R$^2=1-\frac{\ell\left(Z\widehat{\theta}\right)}{\ell\left(\widehat{\theta}_0\right)}=$ 11.1%, where $\ell\left(Z\widehat{\theta}\right)$ is the log-likelihood for the model and $\ell\left(\widehat{\theta}_0\right)$ is the

[9] An efficient algorithm for this Gibbs sampler probit, rbprobitGibbs, is available in the bayesm package of R (http://www.r-project.org/), the open source statistical computing project. Bayesm is a package written to complement Rossi, Allenby, and McCulloch [2005].

[10] See the second appendix to these notes for a brief discussion of *ML* estimation of discrete choice models.

[11] Comparison of estimates based on $n=6$ versus $n=120$ samples produces no difference in *ML* parameter estimates but substantial difference in the *McMC* estimates. The $n=6$ *McMC* estimates are typically larger in absolute value compared to their $n=120$ counterparts. This tends to exagerate heterogeneity in outcomes if we reconnect with the treatment effect examples. The remainder of this discussion focuses on the $n=120$ sample.

log-likelihood with a constant only). However, recall this selection equation works perfectly as a control function in the treatment effect example where high explanatory power does not indicate an adequate model specification (see projections chapter).

Now, we compare the *ML* results with *McMC* data augmentation and the Gibbs sampler probit discussed previously. Statistics from $10,000$ posterior draws following $1,000$ burn-in draws are tabulated below based on the $n = 120$ sample.

| statistic | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ |
|---|---|---|---|---|---|
| mean | $-0.6225$ | $0.5030$ | $-0.1516$ | $-0.7375$ | $0.2310$ |
| median | $-0.6154$ | $0.5003$ | $-0.1493$ | $-0.7336$ | $0.2243$ |
| standard deviation | $0.2189$ | $0.1488$ | $0.2669$ | $0.2057$ | $0.1879$ |
| quantiles: | | | | | |
| 0.025 | $-1.0638$ | $0.2236$ | $-0.6865$ | $-1.1557$ | $-0.1252$ |
| 0.25 | $-0.7661$ | $0.4009$ | $-0.3286$ | $-0.8720$ | $0.1056$ |
| 0.75 | $-0.4713$ | $0.6007$ | $0.02757$ | $-0.5975$ | $-0.3470$ |
| 0.975 | $-0.1252$ | $0.1056$ | $0.2243$ | $0.3549$ | $0.6110$ |

Sample statistics for data augmented Gibbs *McMC* probit posterior draws
$DGP: U_D = Z\theta + \varepsilon, \quad Z = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 \end{bmatrix}$

As expected, the means, medians, and standard errors of the *McMC* probit estimates correspond quite well with *ML* probit estimates.

## 7.11 Logit example

Next, we apply logistic regression (logit for short) to the same ($n = 120$) data set. We compare *MLE* results with two *McMC* strategies: (1) logit estimated via a random walk Metropolis-Hastings (*MH*) algorithm without data augmentation and (2) a uniform data augmented Gibbs sampler logit.

### 7.11.1 Random walk MH for logit

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.

We wish to draw from the posterior

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

where the log likelihood is

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

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.

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 $\theta^0$ 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, $\theta_j$,

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

3. Draw a vector, $\theta^*$ from $N\left(\theta^{k-1}, \sigma^2 I\right)$. Notice, for the random walk, the tuning parameter, $\sigma^2$, is the key. If $\sigma^2$ is chosen too large, then the algorithm will reject the proposal draw frequently and will converge slowly, If $\sigma^2$ 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 $\alpha$ where

$$\alpha = \begin{cases} \min\left(1, \frac{\Pr(\theta^* \mid D, Z) q\left(\theta^*, \theta^{k-1}\right)}{\Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right)}\right) & \Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right) > 0 \\ 1 & \Pr\left(\theta^{k-1} \mid D, Z\right) q\left(\theta^{k-1}, \theta^*\right) = 0 \end{cases}$$

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 $\alpha$ simplifies as $\frac{q\left(\theta^{*},\theta^{k-1}\right)}{q\left(\theta^{k-1},\theta^{*}\right)}$ drops out. Hence, we calculate

$$\alpha = \begin{cases} \min\left(1, \frac{\Pr(\theta^{*}|D,Z)}{\Pr\left(\theta^{k-1}|D,Z\right)}\right) & \Pr\left(\theta^{k-1} \mid D,Z\right) q\left(\theta^{k-1},\theta^{*}\right) > 0 \\ 1 & \Pr\left(\theta^{k-1} \mid D,Z\right) q\left(\theta^{k-1},\theta^{*}\right) = 0 \end{cases}$$

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 $\alpha$ accept the proposal draw, $\theta^{*}$.

6. Repeat $K$ times until the distribution converges.

### *7.11.2 Uniform Gibbs sampler for logit*

On the other hand, the uniform data augmented Gibbs sampler logit specifies a complete set of conditional posteriors developed as follows. Let

$$D_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1-\pi_i \end{cases}, i = 1,2,\ldots,n$$

where $\pi_i = \frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]} = F_V\left(Z_i^T\theta\right)$, or $\log\frac{\pi_i}{1-\pi_i} = Z_i^T\theta$, and $F_V\left(Z_i^T\theta\right)$ is the cumulative distribution function of the logistic random variable $V$. Hence, $\pi_i = \Pr\left(U < \frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]}\right)$ where $U$ has a uniform$(0,1)$ distribution. Then, given the priors for $\theta$, $p(\theta)$, the joint posterior for the latent variable $u = (u_1, u_2, \ldots, u_n)$ and $\theta$ given the data $D$ and $Z$ is

$$\Pr(\theta, u \mid D, Z) \propto p(\theta) \prod_{i=1}^{n} \left\{ \begin{array}{c} I\left(u_i \leq \frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]}\right) I(D_i = 1) \\ +I\left(u_i > \frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]}\right) I(D_i = 0) \end{array} \right\} I(0 \leq u_i \leq 1)$$

where $I(X \in A)$ is an indicator function that equals one if $X \in A$, and zero otherwise.

Thus, the conditional posterior for the latent (uniform) variable $u$ is

$$\Pr(u_i \mid \theta, D, Z) \sim \begin{cases} Uniform\left(0, \frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]}\right) & \text{if } D_i = 1 \\ Uniform\left(\frac{\exp\left[Z_i^T\theta\right]}{1+\exp\left[Z_i^T\theta\right]}, 1\right) & \text{if } D_i = 0 \end{cases}$$

Since the joint posterior can be written

$$\Pr(\theta, u \mid D, Z) \propto p(\theta) \prod_{i=1}^{n} \left\{ \begin{array}{c} I\left(Z_i^T\theta \geq \log\frac{u_i}{1-u_i}\right) I(D_i = 1) \\ +I\left(Z_i^T\theta < \log\frac{u_i}{1-u_i}\right) I(D_i = 0) \end{array} \right\} I(0 \leq u_i \leq 1)$$

we have

$$\textstyle\sum_{j=1}^{5} Z_{ij}\theta_j \geq \log \frac{u_i}{1-u_i} \quad \text{if } D_i = 1$$

so

$$\theta \geq \frac{1}{Z_{ik}} \left( \log \frac{u_i}{1-u_i} - \sum_{j \neq k} Z_{ij}\theta_j \right)$$

for all samples for which $D_i = 1$ and $Z_{ik} > 0$, as well as for all samples for which $D_i = 0$ and $Z_{ik} < 0$. Similarly,

$$\theta_k < \frac{1}{Z_{ik}} \left( \log \frac{u_i}{1-u_i} - \sum_{j \neq k} Z_{ij}\theta_j \right)$$

for all samples for which $D_i$ =1 and $Z_{ik} > 0$, as well as for all samples for which $D_i = 0$ and $Z_{ik} < 0$.[12] Let $A_k$ and $B_k$ be the sets defined by the above, that is,

$$A_k = \{ i : ((D_i = 1) \cap (Z_{ik} > 0)) \cup ((D_i = 0) \cap (Z_{ik} < 0)) \}$$

and

$$B_k = \{ i : ((D_i = 0) \cap (Z_{ik} > 0)) \cup ((D_i = 1) \cap (Z_{ik} < 0)) \}$$

A diffuse prior $p(\theta) \propto 1$ combined with the above gives the conditional posterior for $\theta_k$, $k = 1, 2, \ldots, 5$, given the other $\theta$'s and latent variable, $u$.

$$p(\theta_k \mid \theta_{-k}, u, D, Z) \sim Uniform(a_k, b_k)$$

where $\theta_{-k}$ is a vector of parameters except $\theta_k$,

$$a_k = \max_{i \in A_k} \left[ \frac{1}{Z_{ik}} \left( \log \frac{u_i}{1-u_i} - \sum_{j \neq k} Z_{ij}\theta_j \right) \right]$$

and

$$b_k = \min_{i \in B_k} \left[ \frac{1}{Z_{ik}} \left( \log \frac{u_i}{1-u_i} - \sum_{j \neq k} Z_{ij}\theta_j \right) \right]$$

The Gibbs sampler is implemented by drawing $n$ values of $u$ in one block conditional on $\theta$ and the data, $D$, $Z$. The elements of $\theta$ are drawn successively, each conditional on $u$, the remaining parameters, $\theta_{-k}$, and the data, $D$, $Z$.

[12] If $Z_{ik} = 0$, the observation is ignored as $\theta_k$ is determined by the other regressor values.

### *7.11.3 Comparison of logit results*

*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 | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ |
|---|---|---|---|---|---|
| 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
$DGP : U_D = Z\theta + \varepsilon, \quad Z = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 \end{bmatrix}$

Statistics from $10,000$ posterior data augmented uniform Gibbs draws following $40,000$ burn-in draws[13] are tabulated below based on the $n = 120$ sample.

| statistic | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ |
|---|---|---|---|---|---|
| mean | $-1.015$ | $0.8259$ | $-0.3375$ | $-1.199$ | $0.3529$ |
| median | $-1.011$ | $0.8126$ | $-0.3416$ | $-1.2053$ | $0.3445$ |
| standard deviation | $0.3014$ | $0.2039$ | $0.3748$ | $0.2882$ | $0.2884$ |
| quantiles: | | | | | |
| 0.025 | $-1.6399$ | $0.3835$ | $-1.1800$ | $-1.9028$ | $-0.2889$ |
| 0.25 | $-1.2024$ | $0.6916$ | $-0.5902$ | $-1.3867$ | $0.1579$ |
| 0.75 | $-0.8165$ | $0.9514$ | $-0.0891$ | $-1.0099$ | $0.5451$ |
| 0.975 | $-0.4423$ | $1.2494$ | $0.3849$ | $-0.6253$ | $0.9451$ |

Sample statistics for *uniform Gibbs McMC* logit posterior draws
$DGP : U_D = Z\theta + \varepsilon, \quad Z = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 \end{bmatrix}$

[13] Convergence to marginal posterior draws is much slower with this algorithm.

As expected, the means, medians, and standard errors of the *McMC* logit estimates correspond well with each other and the *ML* logit estimates. Now that we've developed *McMC* data augmentation for the choice or selection equation, we return to the discussion of causal effects (initiated in the classical linear models chapter) and discuss data augmentation for the counterfactuals as well as latent utility.