# Notes on Bayesian Computation and MCMC

Adam N. Smith
July 31, 2021

*This note reviews some of the key results underlying MCMC theory, discusses the theoretical underpinnings of popular MCMC algorithms (the Metropolis-Hastings algorithm and Gibbs sampler), and presents a few applications in the context of economic choice models.*

## 1 Preliminaries

Markov chain Monte Carlo (MCMC) methods are an indispensable tool in the Bayesian paradigm. In some sense, MCMC put Bayesian analysis "on the map" by making it feasible to generate posterior samples from a much wider class of Bayesian models. While non-conjugate priors and normalizing constants would pose challenges to analytical posterior sampling solutions (or preclude them altogether), they are no longer issues with MCMC. The broad idea of MCMC is to construct a Markov chain that converges to the required posterior distribution. The goal of this note is to define canonical MCMC algorithms and briefly explain why they work.

It is first worth mentioning the fascinating history of Bayesian computation and MCMC methods, which has been discussed at length in Robert and Casella (2011) and Martin et al. (2020). A short version is as follows. The underlying methods date back to Los Alamos, New Mexico and statistical physics applications during World War II. The earliest developments include the Monte Carlo method (Metropolis and Ulam, 1949) and the Metropolis algorithm (Metropolis et al., 1953), which was then extended by Hastings (1970). The ideas of Gibbs sampling also have roots in Hastings (1970), but were rigorously laid out in Geman and Geman (1984). A seminal paper by Gelfand and Smith (1990) popularized MCMC to the Bayesian community, which sparked an MCMC revolution in the '90s. Robert and Casella (2011) also cites a conference held at Ohio State University in February 1991 as a seed of this revolution. The conference was organized by Alan Gelfand, Prem Goel, and Adrian Smith and included talks from nearly everyone who would now make the "Who's Who of MCMC" list. The conference program is included in the appendix of Robert and Casella (2011).

**Definitions** A **Markov chain** is a sequence of random variables $\theta^{(0)}, \theta^{(1)}, \ldots$ evolving over time, where this evolution adherers to a specific Markov structure. That is, the present state $\theta^{(r)}$ only depends on the past through the last state $\theta^{(r-1)}$ or, formally, $P(\theta^{(r)}|\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(r-1)}) = P(\theta^{(r)}|\theta^{(r-1)})$. A Markov chain is governed by the **transition kernel** $P(\theta, A)$ which defines the probability of reaching set $A \subset \Theta$ from state $\theta \in \Theta$. Its density $p(\theta, \cdot)$ is defined as $P(\theta, A) = \int_A p(\theta, \vartheta) d\vartheta$ and is a valid pdf *conditional on* $\theta$.

The broad goal of Markov chain theory is to identify conditions under which a Markov chain possesses a (unique) stationary distribution and also converges to this stationary distribution in the limit. Formally, the distribution $\pi$ is a **stationary distribution** of the Markov transition kernel $P$ if

$$\int_\Theta P(\theta, A)\pi(\theta)d\theta = \int_A \pi(\vartheta)d\vartheta \tag{1}$$

To understand why MCMC algorithms "work", we first need to understand some of the key properties of Markov chain theory which will help us establish the existence and uniqueness of, and convergence to, this stationary distribution.

1. (Irreducibility) A Markov chain is **$\pi$-irreducible** if for any initial state $\theta$, $P(\theta, A) > 0$ whenever $\pi(A) > 0$. That is, there is always some way to reach any state (with positive probability under $\pi$) from any other state.

2. (Recurrence) A stronger version of irreducibility is Harris recurrence. A Markov chain is **Harris recurrent** if, for all $A$ with $\pi(A) > 0$ and all $\theta$, the chain will reach $A$ with probability 1. As Chan and Geyer (1994) put it, "Harris recurrence essentially says that there is no measure-theoretic pathology ... the main point of Harris recurrence is that asymptotics do not depend on the starting distribution."

3. (Reversibility) A Markov chain is **reversible** if the distribution of $(\theta^{(r)}, \theta^{(r+1)})$ is the same as the distribution of $(\theta^{(r+1)}, \theta^{(r)})$. This imposes exchangeability between $\theta^{(r)}$ and $\theta^{(r+1)}$.

4. (Detailed Balance) A Markov chain with transition kernel $P$ satisfies **detailed balance** with respect to the distribution $\pi$ if

$$P(\theta, \vartheta)\pi(\theta) = P(\vartheta, \theta)\pi(\vartheta) \qquad \text{for all } \theta, \vartheta \in \Theta. \tag{2}$$

Importantly, detailed balance is a sufficient (but not necessary) condition for reversibility.

5. (Periodicity) A Markov chain is **periodic** if there are portions of the state space that it can only visit at certain regularly spaced times; otherwise, the chain is **aperiodic**. A sufficient (but not necessary) condition for an aperiodic chain is $P(\theta^{(r)} = \theta^{(r-1)}) > 0$.

6. (Ergodicity) A Markov chain is **ergodic** if its limiting distribution equals its stationary distribution. That is, if for all sets $A$,

$$\lim_{r\to\infty} \left\| P^{(r)}(\theta, A) - \pi(A) \right\| = 0 \tag{3}$$

where $\|\cdot\|$ is the total variation norm.

**Key Results** We now state a few key results connecting the properties above to desired properties of Markov chains. A formal presentation of these results and discussion of their proofs can be found in Robert and Casella (2004).

R1 If the Markov transition kernel $P$ is **reversible** w.r.t. $\pi$, then $\pi$ is a stationary distribution of $P$.

R2 If the Markov transition kernel $P$ is **$\pi$-irreducible**, then $\pi$ is the unique stationary distribution of $P$.

R3 If the Markov transition kernel $P$ is **$\pi$-irreducible** and **aperiodic**, then $\|P^{(r)}(\theta, A) - \pi(A)\| \to 0$ for almost-every $\theta$. If $P$ is also **Harris recurrent**, then convergence holds for every $\theta$.

A corollary to R3 is an ergodic theorem (i.e., a law of large numbers) which says that sample averages converge to the appropriate population integral.

$$\lim_{R\to\infty} \frac{1}{R} \sum_{r=1}^{R} g(\theta^{(r)}) = \mathbb{E}^{\pi}[g(\theta)] \tag{4}$$

This last result constitutes the second "MC" of MCMC. After we construct the appropriate Markov chain, we can use Monte Carlo integration to compute posterior moments for any set of parameters (or functions of parameters).

## 2 MCMC Algorithms

We now turn to the construction of MCMC algorithms. Specifically, an MCMC algorithm for simulating a distribution $\pi$ is any algorithm producing an ergodic Markov chain $\{\theta^{(r)}\}$ whose stationary distribution is $\pi$. In our context, we need $\pi$ to be a posterior $\pi(\theta|x)$ so that $\{\theta^{(r)}\}$ can be treated as draws from $\pi(\theta|x)$. Our "recipe" to check the validity of an MCMC algorithm is to verify that the transition kernel: (i) satisfies detailed balance with respect to $\pi$; (ii) is $\pi$-irreducible; and (iii) is aperiodic. Together, (i) and (ii) show that the Markov chain has the posterior as its unique stationary distribution, and (ii) and (iii) show that the chain is ergodic and thus converges to the posterior.

### 2.1 Metropolis-Hastings Algorithm

One of the most general and widely-used MCMC algorithms is the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). Given a target posterior $\pi(\theta|x)$ and conditional density $q(\vartheta|\theta)$, the algorithm generates a chain $\{\theta^{(r)}\}$ through the following steps.

**Algorithm 1: Metropolis-Hastings**

Initialize $\theta^{(0)}$
For $r = 1, \ldots, R$

1. Let $\theta = \theta^{(r-1)}$ and generate a candidate value $\vartheta \sim q(\cdot|\theta)$
2. Compute the MH acceptance probability
$$\alpha(\theta, \vartheta) = \min\left\{1, \frac{\pi(\vartheta|x)}{\pi(\theta|x)}\frac{q(\theta|\vartheta)}{q(\vartheta|\theta)}\right\}$$
3. Set $\theta^{(r)} = \vartheta$ with probability $\alpha(\theta, \vartheta)$
Otherwise set $\theta^{(r)} = \theta$

**Proof** Our first task is to verify that the MH algorithm is valid (formal results can be found in Tierney, 1994). Let $\theta$ and $\vartheta$ two states of the chain.

(i) (Detailed Balance) The MH transition kernel is defined as

$$P(\theta, \vartheta) = \alpha(\theta, \vartheta) q(\vartheta|\theta)$$

First step is to show that $P(\theta, \vartheta)$ satisfies the detailed balance condition.

$$\alpha(\theta, \vartheta) q(\vartheta|\theta) \pi(\theta|x) = \alpha(\vartheta, \theta) q(\theta|\vartheta) \pi(\vartheta|x)$$

First assume $\pi(\vartheta|x)q(\theta|\vartheta) > \pi(\theta|x)q(\vartheta|\theta)$ so that $\alpha(\theta, \vartheta) = 1$ and, by symmetry, $\alpha(\vartheta, \theta) = \pi(\theta|x)q(\vartheta|\theta)/(\pi(\vartheta|x)q(\theta|\vartheta))$. It follows that

$$\begin{aligned} &\text{(LHS)} \quad \alpha(\theta, \vartheta) q(\vartheta|\theta) \pi(\theta|x) = q(\vartheta|\theta)\pi(\theta|x) \\ &\text{(RHS)} \quad \alpha(\vartheta, \theta) q(\theta|\vartheta) \pi(\vartheta|x) = \frac{\pi(\theta|x)q(\vartheta|\theta)}{\pi(\vartheta|x)q(\theta|\vartheta)} q(\theta|\vartheta)\pi(\vartheta|x) = \pi(\theta|x)q(\vartheta|\theta) \end{aligned}$$

so the LHS and RHS are equal and detailed balance is satisfied. Note that because of symmetry, the same is true for the case when $\pi(\vartheta|x)q(\theta|\vartheta) \leq \pi(\theta|x)q(\vartheta|\theta)$. Therefore, the MH update is reversible.

(ii) ($\pi$-Irreducible) A sufficient condition for irreducibility is that $q(\vartheta|\theta) > 0$ for all $(\vartheta, \theta)$ in the support of the posterior $\pi(\theta|x)$. Moreover, by Corollary 2 in Tierney (1994), if a Metropolis-Hastings chain is $\pi$-irreducible then it is Harris recurrent.

(iii) (Aperiodic) Since the MH acceptance probability guarantees that some candidate draws will be "rejected" then $P(\theta^{(r)} = \theta^{(r-1)}) > 0$ and the chain is aperiodic.

By (ii) and (iii) above, the MH chain is ergodic as desired!

**Virtues** A few comments are also in order about the virtues of the MH algorithm. First, the algorithm is both simple and remarkably general in the sense that it "works" under very mild assumptions on proposal densities and with no assumptions of likelihood structure or prior conjugacy. Moreover, the ratio of posteriors in the MH acceptance ratio implies that the normalizing constants cancel, and so the algorithm only requires evaluations of the likelihood and prior. This is one of the main reasons why the MH algorithm was a breakthrough for Bayesian computation.

**Proposal Densities** There are two main classes of proposal densities, each leading to a different "type" of MH algorithm. The first specifies proposals of the form:

$$q(\vartheta|\theta) = q(\vartheta) \tag{5}$$

and so the proposed value is independent of the current state. MH algorithms with such proposals are referred to as *independence MH algorithms*. The second specifies proposals of the form:

$$q(\vartheta|\theta) = q(\|\vartheta - \theta\|) \tag{6}$$

and are thus symmetric. In this case, we can we can write $\vartheta = \theta + \epsilon$ where $\epsilon$ has a symmetric distribution. MH algorithms with this type of symmetric proposal density are referred to as *random-walk MH algorithms*. One feature of symmetric proposals is that the ratio of proposals will cancel in the MH acceptance ratio, and so the only objects that must be evaluated are the likelihood and prior. In fact, the original Metropolis algorithm (Metropolis et al., 1953) assumed symmetric proposals and so one innovation of Hastings (1970) was to provide an extension to asymmetric proposals.

## 2.2 Gibbs Sampler

The MH is very general in the sense that the only requirements are evaluations of the likelihood, prior, and proposal densities. The Gibbs sampler requires a stronger set of conditions to be met – namely, given a $p$-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_p)$ and target posterior $\pi(\theta|x)$, we must be able to sample from each parameter's **full conditional distribution** $\pi(\theta_j|\theta_{-j}, x)$. Consequently, Gibbs samplers are most commonly used when some form of *conditional conjugacy* exists so that the full conditionals can be expressed analytically.

The idea of Gibbs sampling is in some ways both rooted in and justified by the **Hammersley-Clifford Theorem**, which proves that we can write out a joint distribution $p(\theta_1, ..., \theta_p)$ in terms of only the full conditional distributions $p(\theta_j|\theta_{-j})$. For example, consider the bivariate distribution $p(\theta_1, \theta_2)$. It follows that

$$
\begin{aligned}
p(\theta_1, \theta_2) &= p(\theta_2|\theta_1) \times p(\theta_1) \\
&= p(\theta_2|\theta_1) \times \frac{1}{\frac{1}{p(\theta_1)}} \\
&= p(\theta_2|\theta_1) \times \frac{1}{\int \frac{p(\theta_2)}{p(\theta_1)} d\theta_2} \\
&= p(\theta_2|\theta_1) \times \frac{1}{\int \frac{p(\theta_1,\theta_2)/p(\theta_1)}{p(\theta_1,\theta_2)/p(\theta_2)} d\theta_2} \\
&= p(\theta_2|\theta_1) \times \frac{1}{\int \frac{p(\theta_2|\theta_1)}{p(\theta_1|\theta_2)} d\theta_2}
\end{aligned} \tag{7}
$$

and so the set of full conditional distributions, $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$, summarize all information in the joint distribution.

Formally, given a target posterior $\pi(\theta|x)$, the Gibbs sampler generates a chain $\{\theta^{(r)}\}$ by iteratively sampling from each parameter's full conditional distribution.

**Algorithm 2: Gibbs Sampler**

Initialize $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$
For $r = 1, \ldots, R$

1. Sample $\theta_1^{(r)} \sim \pi(\theta_1|\theta_1^{(r-1)}, ..., \theta_p^{(r-1)}, x)$
2. Sample $\theta_2^{(r)} \sim \pi(\theta_2|\theta_1^{(r)}, \theta_3^{(r-1)}, ..., \theta_p^{(r-1)}, x)$

$\vdots$

$p$. Sample $\theta_p^{(r)} \sim \pi(\theta_p|\theta_1^{(r)}, \theta_2^{(r)}, ..., \theta_{p-1}^{(r)}, x)$

**Proof** To demonstrate validity, we will show that the Gibbs sampler is a special case of the MH algorithm with acceptance probability equal to 1. Consider a move from $\theta = (\theta_1, \ldots, \theta_j, \ldots, \theta_p) \to \vartheta = (\theta_1, \ldots, \vartheta_j, \ldots, \theta_p)$. That is, we will modify the $j$th component of the $\theta$ vector. The proposals under a Gibbs sampler take the form:

$$
\begin{aligned}
q(\vartheta|\theta) &= \pi(\vartheta_j|\theta_{-j}, x) \\
q(\theta|\vartheta) &= \pi(\theta_j|\theta_{-j}, x)
\end{aligned}
$$

and so the MH acceptance ratio is equal to:

$$\frac{\pi(\vartheta|x)}{\pi(\theta|x)}\frac{q(\theta|\vartheta)}{q(\vartheta|\theta)} = \frac{\pi(\vartheta|x)}{\pi(\theta|x)}\frac{\pi(\theta_j|\theta_{-j},x)}{\pi(\vartheta_j|\theta_{-j},x)} = \frac{\pi(\vartheta_j|\theta_{-j},x)\pi(\theta_{-j}|x)}{\pi(\theta_j|\theta_{-j},x)\pi(\theta_{-j}|x)}\frac{\pi(\theta_j|\theta_{-j},x)}{\pi(\vartheta_j|\theta_{-j},x)} = 1.$$

Thus, a Gibbs update for $\theta_j$ is equivalent to an MH update where the acceptance probability is equal to 1.

# 3 Examples

In this section, we show how the MCMC algorithms discussed above can be applied to Bayesian choice models. We specifically consider a binary probit model (Gibbs sampler), a multinomial logit model (MH algorithm), and a hierarchical multinomial logit model (hybrid Gibbs or Metropolis-within-Gibbs sampler). These examples can also be found in Chapters 3.8, 3.11, and 5.4 of Rossi et al. (2005).

## 3.1 Binary probit model (Gibbs sampler)

The binary probit model is a latent variable model with a binary outcome $y_i \in \{0, 1\}$.

$$z_i = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \tag{8}$$

$$y_i = \begin{cases} 1 & \text{if } z_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The first step to a Bayesian analysis is to write down the likelihood, prior, and posterior.

(i) Likelihood

$$p(y|X,\beta) = \prod_{i=1}^{n} \Phi(x_i'\beta)^{y_i}[1-\Phi(x_i'\beta)]^{1-y_i}$$

(ii) Prior

$$\beta \sim N(\bar{\beta}, A^{-1})$$

(iii) Posterior

$$\begin{aligned}\pi(\beta|y,x) &\propto p(y|X,\beta)p(\beta) \\ &\propto \left(\prod_{i=1}^{n} p(y_i|x_i,\beta)\right) p(\beta) \\ &\propto \left(\prod_{i=1}^{n} \Phi(x_i'\beta)^{y_i}[1-\Phi(x_i'\beta)]^{1-y_i}\right) \exp\left(-\frac{1}{2}(\beta-\bar{\beta})'A(\beta-\bar{\beta})\right)\end{aligned}$$

The computational challenge with this model is that the posterior does not have an analytic expression (conjugate priors do not exist). Direct analytic sampling from the posterior is therefore infeasible.

Let's see how we can set up a Gibbs sampler to solve this problem. The first "solution" is to treat the latent $z_i$ as a model parameter, thus augmenting the parameter vector to be $(\beta, z_1, \ldots, z_n)$. This is referred to as **data augmentation** and was a contemporaneous development to the Gibbs sampler (Tanner and Wong, 1987). We now have a set of two full conditional distributions: $\pi(\beta|z_1,\ldots,z_n,y,X)$ and $\pi(z_1,\ldots,z_n|\beta,y,X)$.

Note that conditional on the $z$ vector, the full conditional for $\beta$ becomes the posterior of a Bayesian linear regression model with normal conjugate priors (or more specifically, *conditionally* conjugate priors). Therefore, we have

$$\beta|z,y,X \sim N(\tilde{\beta},(X'X+A)^{-1}) \tag{10}$$

where $\tilde{\beta} = (X'X+A)^{-1}(X'z+A\bar{\beta})$. So conditional on knowing the latent outcome variables, we have an analytic expression for the posterior from which we can easily generate samples. Turning to the second full conditional, $z_i$ is a truncated normal random variable with truncation points governed by the choice outcome $y_i$. That is,

$$\begin{aligned} z_i|\beta,y_i=1,x_i &\sim N(x_i'\beta,1)\cdot I_{[0,\infty)}(z_i) \\ z_i|\beta,y_i=0,x_i &\sim N(x_i'\beta,1)\cdot I_{(-\infty,0)}(z_i)\end{aligned} \tag{11}$$

and so $z_i$ is truncated below at 0 if $y_i = 1$ and truncated above at 0 if $y_i = 0$. Sampling from this full conditional can be made using a function like `rtnorm()` in R, for example.

Together, (10) and (11) give us analytic expressions for the necessary full condi-

tional distributions. We can then use a Gibbs sampler to sample from the posterior.[1] Specifically, initialize $\beta^{(0)}, z_1^{(0)}, \ldots, z_n^{(0)}$ and then for $r = 1, \ldots, R$, do the following.

- Sample $\beta^{(r)}|z_1^{(r-1)}, \ldots, z_n^{(r-1)}, y, X$ from (10)
- Sample $z_1^{(r)}|\beta^{(r)}, y, X$ from (11)

  $\vdots$

- Sample $z_n^{(r)}|\beta^{(r)}, y, X$ from (11)

Note that the terms $z_{-i}^{(r-1)}$ are absent from the right-hand-size of the $z_i$ full conditional since the $z_i$'s are assumed to be iid.

### 3.2 Multinomial logit model (Metropolis-Hastings algorithm)

The multinomial logit model is another latent variable model, but now with a multinomial outcome $y_i \in \{1, \ldots, J\}$ where

$$u_{ij} = x_{ij}'\beta + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{TIEV}$$

and

$$y_i = j \quad \text{if} \quad u_{ij} \geq u_{ik} \text{ for all } k \neq j.$$

We again start by writing down the likelihood, prior, and posterior.

(i) Likelihood

$$p(y|\beta, X) = \prod_{i=1}^{n} \prod_{j=1}^{J} p_{ij}^{\mathbf{1}(y_i=j)}$$
$$p_{ij} = \frac{\exp(x_{ij}'\beta)}{\sum_{k=1}^{J} \exp(x_{ik}'\beta)}$$

(ii) Prior

$$\beta \sim N(\bar{\beta}, A^{-1})$$

[1] The function `rbprobitGibbs()` in the `bayesm` R package (Rossi, 2019) implements this Gibbs sampling algorithm for the binary probit model.

(iii) Posterior

$$\begin{aligned}\pi(\beta|y,X) &\propto p(y|X,\beta)p(\beta) \\ &\propto \left(\prod_{i=1}^{n} p(y_i|x_i,\beta)\right) p(\beta) \\ &\propto \left(\prod_{i=1}^{n}\prod_{j=1}^{J}\left[\frac{\exp(x_{ij}'\beta)}{\sum_{k=1}^{J}\exp(x_{ik}'\beta)}\right]^{\mathbf{1}(y_i=j)}\right)\exp\left(-\frac{1}{2}(\beta-\bar{\beta})'A(\beta-\bar{\beta})\right)\end{aligned}$$

Just like the case of the binary probit model, the computational challenge with the multinomial logit model is that the posterior does not have an analytic expression and so direct analytic sampling from the posterior above is infeasible. Moreover, we cannot find conjugate or even conditionally conjugate priors for $\beta$, so a Gibbs sampler is also infeasible.

We therefore turn to a random-walk MH algorithm.[2] Initialize $\beta^{(0)}$ and then for $r = 1, \ldots, R$, do the following.

- Generate a candidate value $\beta^* \sim N(\beta^{(r-1)}, s^2)$
- Compute the MH acceptance probability

$$\alpha(\beta^{(r-1)}, \beta^*) = \min\left\{1, \frac{p(y|X,\beta^*)p(\beta^*)}{p(y|X,\beta^{(r-1)})p(\beta^{(r-1)})}\right\}$$

- Set $\beta^{(r)} = \beta^*$ with probability $\alpha(\beta^{(r-1)}, \beta^*)$ and set $\beta^{(r)} = \beta^{(r-1)}$ otherwise

Note that the proposal density is parameterized by a variance term $s^2$ which acts as a "step size" of the MH proposal. For guidance on choosing $s^2$ in the context of logit models, see Section 3.11 of Rossi et al. (2005). Also note that the random-walk proposal is symmetric and so the MH acceptance ratio only involves evaluations of the likelihood and prior (and not the proposal).

[2] The function `rmnlIndepMetrop()` in the `bayesm` R package (Rossi, 2019) implements an independence MH algorithm for the multinomial logit model.

### 3.3 Hierarchical multinomial logit model (Metropolis-within-Gibbs)

A hierarchical multinomial logit model extends the logit model above to a panel data setting with customers $i = 1, \ldots, n$ and purchase occasions $t = 1, \ldots, T_i$:

$$u_{ijt} = x'_{ijt}\beta_i + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim \text{TIEV}$$

and

$$y_{it} = j \quad \text{if} \quad u_{ijt} \geq u_{ikt} \text{ for all } k \neq j.$$

We again start by writing down the likelihood, prior, and posterior.

(i) Likelihood

$$p(y|X, \beta) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} \prod_{j=1}^{J} p_{ijt}^{\mathbf{1}(y_{it}=j)}$$
$$p_{ijt} = \frac{\exp(x'_{ijt}\beta_i)}{\sum_{k=1}^{J} \exp(x'_{ikt}\beta_i)}$$

(ii) Prior

$$\beta_i|\Delta, V_\beta \sim N(\Delta' z_i, V_\beta)$$

Here $z_i$ is a $d$-vector of observed customer characteristics variables (like demographics) and $\Delta$ is a $d \times p$ matrix describing how average preferences change with observed characteristics. While the mean function captures "observed heterogeneity", the covariance matrix $V_\beta$ captures "unobserved heterogeneity." That is, two customers are allowed to have different $\beta_i$ vectors even if their $z_i$ vectors are identical. Since we want to make inferences about $(\Delta, V_\beta)$, we need to specify a second-stage prior:

$$\text{vec}(\Delta)|V_\beta \sim N(\text{vec}(\bar{\Delta}), V_\beta \otimes A^{-1})$$
$$V_\beta \sim IW(\nu, V)$$

which is the conjugate prior for a multivariate regression model (Rossi et al., 2005).

(iii) Posterior

$$
\begin{aligned}
&\pi(\beta, \Delta, V_\beta | y, X, z) \\
&\quad \propto p(y|X, \beta) p(\beta|\Delta, V_\beta) p(\Delta, V_\beta) \\
&\quad \propto \left( \prod_{i=1}^{n} \left[ \prod_{t=1}^{T_i} p(y_{it}|x_i, \beta_i) \right] p(\beta_i|\Delta, V_\beta) \right) p(\Delta|V_\beta) p(V_\beta) \\
&\quad \propto \left( \prod_{i=1}^{n} \left[ \prod_{t=1}^{T_i} \prod_{j=1}^{J} \left[ \frac{\exp(x_{ij}'\beta)}{\sum_{k=1}^{J} \exp(x_{ik}'\beta)} \right]^{\mathbf{1}(y_{it}=j)} \right] \exp\left( -\frac{1}{2}(\beta_i - \Delta' z_i)' V_\beta^{-1} (\beta_i - \Delta' z_i) \right) \right) \\
&\qquad \times |V_\beta|^{-d/2} \exp\left( -\frac{1}{2} \mathrm{tr}\left( (\Delta - \bar{\Delta})' A (\Delta - \bar{\Delta}) V_\beta^{-1} \right) \right) \\
&\qquad \times |V_\beta|^{-(\nu+p+1)/2} \exp\left( -\frac{1}{2} \mathrm{tr}\left( V V_\beta^{-1} \right) \right)
\end{aligned}
$$

Just like both models described above, the computational challenge with this model is that the posterior does not have an analytic expression and so direct analytic sampling from the posterior is infeasible. But the hierarchical nature of the model allows us to sample "blocks" of parameters at a time. In particular, we can design a Gibbs sampler that will iteratively sample from the following full conditionals:

$$
\begin{aligned}
&\beta | \Delta, V_\beta, y, X \\
&\Delta, V_\beta | \beta, y, X
\end{aligned}
$$

where the first block is the set of customer-level parameters $\beta_1, \ldots, \beta_n$ and the second is the set of population-level parameters $(\Delta, V_\beta)$. Note that the sampling problem associated with the $\beta$ vector is the same problem that appeared above for the multinomial logit model. The only difference is that here we have panel data, and so we must loop over customers and do a total of $n$ MH updates. Then conditional on the $\beta$'s, there is an analytic expression for the full conditional for $(\Delta, V_\beta)$ and so sampling is exact.

We will use a Metropolis-within-Gibbs algorithm to sample from the posterior above.[3] This hybrid algorithm gets its name from the fact we are iteratively sampling between full conditionals (i.e., Gibbs sampling), but we are swapping in an

[3]The function `rhierMnlRwMixture()` in the `bayesm` R package (Rossi, 2019) implements this Metropolis-within-Gibbs algorithm for the hierarchical multinomial logit model. The only difference is that the function is more general and allows for mixtures of normals heterogeneity.

MH update for parameters with non-conjugate priors and complicated full conditionals (i.e., the $\beta_i$'s). Specifically, initialize the vector of model parameters $\beta_1^{(0)}, \ldots, \beta_1^{(0)}, \Delta^{(0)}, V_\beta^{(0)}$ and then for $r = 1, \ldots, R$, do the following.

- (MH update) For $i = 1, \ldots, n$
  - Generate a candidate value $\beta_i^* \sim N(\beta_i^{(r-1)}, s^2 V_\beta)$
  - Compute the MH acceptance probability

$$\alpha(\beta_i^{(r-1)}, \beta_i^*) = \min \left\{ 1, \frac{p(y_i|x_i, \beta_i^*) p(\beta_i^*|\Delta^{(r-1)}, V_\beta^{(r-1)})}{p(y_i|x_i, \beta_i^{(r-1)}) p(\beta^{(r-1)}|\Delta^{(r-1)}, V_\beta^{(r-1)})} \right\}$$

  - Set $\beta_i^{(r)} = \beta^*$ with probability $\alpha(\beta_i^{(r-1)}, \beta^*)$ and set $\beta_i^{(r)} = \beta_i^{(r-1)}$ otherwise
- (Gibbs update) Sample $(\Delta^{(r)}, V_\beta^{(r)})$ from the posterior of a Bayesian multivariate regression model. Although the structure of this posterior closely matches that of a Bayesian multiple regression model, there is more matrix algebra given the multivariate nature of the response. See Section 2.12 of Rossi et al. (2005) for details.

## References


Chan, K. S. and Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1747–1758.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2020). Computing Bayes: Bayesian computation from 1763 to the 21st century. *Working Paper*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335—341.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Rossi, P. E. (2019). *bayesm: Bayesian Inference for Marketing/Micro-Econometrics*, R package version 3.1-4 edition.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.