# บทที่ 2
# เอกสารและงานวิจัยที่เกี่ยวข้อง

โดยทั่วไปแล้วเวลาเก็บข้อมูลเพื่อนำมาวิเคราะห์ ข้อมูลที่ได้มามักมีตัวแปรหลายประเภท ดังนั้นในการจัดอันดับโดยใช้การวิเคราะห์ปัจจัย ได้มีการปรับให้สามารถวิเคราะห์ข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ โดยทำการวิเคราะห์ปัจจัยผ่านตัวแปรแฝง $x^*$ แต่ด้วยความซับซ้อนของพารามิเตอร์ทั้งหมดที่ต้องประมาณค่า รวมทั้งตัวแปรแฝงที่ต้องใช้ในการวิเคราะห์ จึงใช้การสุ่มตัวอย่างด้วยวิธีกิบส์และการสุ่มตัวอย่างด้วยวิธีฮิตแอนด์รันที่อยู่ในกลุ่มวิธีลูกโซ่มาร์คอฟมอนติคาร์โล (Markov chain Monte Carlo) หรือ MCMC ในการจำลองข้อมูล ซึ่งลูกโซ่มาร์คอฟจะลู่เข้าสู่การแจกแจงคงตัว (stationary distribution) ภายใต้เงื่อนไขบางประการ แล้วใช้วิธีค่าเฉลี่ยกลุ่ม (Batch mean method) สำหรับการประมาณค่าพารามิเตอร์ และค่าคลาดเคลื่อนมาตรฐาน (SE) ของค่าประมาณ เพื่อนำค่าคลาดเคลื่อนมาตรฐานจากวิธี MCMC ทั้ง 2 มาเปรียบเทียบประสิทธิภาพในการลู่เข้าของตัวประมาณ

## 1. ตัวแบบการวิเคราะห์ปัจจัยสำหรับข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ

ให้ $X$ เป็นเมทริกซ์ของค่าสังเกตขนาด $N \times J$ ซึ่งแต่ละตัวแปรสามารถเป็นได้ทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับที่มี $C_j > 1$ กลุ่ม โดยให้ $i = 1,2,\ldots,N$ แทนดัชนีของหน่วยที่ให้ค่าสังเกต $j = 1,2,\ldots,J$ แทนดัชนีของตัวแปร

สมมติให้ ค่าของแต่ละสมาชิกใน $X$ ได้มาจาก $X^*$ ซึ่งเป็นเมทริกซ์ของตัวแปรแฝง (Latent variable) ขนาด $N \times J$ และจุดตัด $\gamma$

$$x_{ij} = \begin{cases} x_{ij}^* & \text{ถ้า ตัวแปรที่ } j \text{ เป็นตัวแปรแบบต่อเนื่อง} \\ c & \text{ถ้า } x_{ij}^* \in (\gamma_{j(c-1)}, \gamma_{jc}] \text{ และตัวแปรที่ } j \text{ เป็นตัวแปรเชิงอันดับ} \end{cases}$$

โดยที่ $c$ มีค่าเป็นไปได้ตั้งแต่ $1,2,\ldots,C_j$ และให้ $\gamma_{j0} = -\infty, \gamma_{j1} = 0$ และ $\gamma_{jC_j} = \infty$

รูปแบบของการรวมตัวระหว่างตัวแปรสังเกตได้ $X$ ถูกจำลองโดยตัวแบบการวิเคราะห์ปัจจัยผ่านตัวแปรแฝง $X^*$

$$x_i^* = \Lambda \phi_i + \varepsilon_i \quad , i = 1,\ldots,N$$

โดยที่ $\phi_i \overset{iid}{\sim} \mathrm{N}(0,\mathrm{I}) \quad , i = 1,\ldots,N$

และ $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0,\Psi) \quad , i = 1,\ldots,N$

ซึ่งสามารถเขียนให้อยู่ในรูปของ $N$ สมการได้ดังนี้

$$X^*_{N\times J} = \Phi_{N\times 2}\Lambda'_{2\times J} + \mathrm{E}_{N\times J}$$

เมื่อ $X^*$ เป็นเมทริกซ์ของตัวแปรแฝงขนาด $N \times J$ $\Lambda$ เป็นเมทริกซ์ขนาด $J \times 2$ ของสัมประสิทธิ์ของปัจจัย เมื่อกำหนดให้มี 1 ปัจจัย, $\Phi$ เป็นเมทริกซ์ขนาด $N \times 2$ โดยสมาชิกในหลักที่ 1 เป็น 1 ทั้งหมดและสมาชิกในหลักที่ 2 เป็นคะแนนของปัจจัย และ $\Psi$ เป็นเมทริกซ์ทแยงมุมที่มีขนาด $J \times J$ โดยให้ $\Psi_{jj}$ เท่ากับ 1 เมื่อตัวแปรที่ $j$ เป็นตัวแปรอันดับ และ $\Psi_{jj} \overset{iid}{\sim} IG(a_o/2, b_o/2)$ เมื่อตัวแปรที่ $j$ เป็นตัวแปรแบบต่อเนื่อง

ในการวิเคราะห์ปัจจัยเชิงเบส์สำหรับการแจกแจงก่อน (Prior distribution) ของพารามิเตอร์ที่เกี่ยวข้องเป็นดังต่อไปนี้

$\Lambda$ จะต้องมีสมาชิกอย่างน้อยหนึ่งตัวในหลักที่ 2 ที่ถูกกำหนดให้มีเครื่องหมายเป็นบวกหรือลบเท่านั้น เพื่อกำจัดปัญหาการไม่แปรเปลี่ยนภายใต้การหมุน (Rotational invariance) โดยสมมติให้สมาชิกของ $\Lambda$ ที่ถูกกำหนดเครื่องหมาย มีการแจกแจงปกติแบบตัดปลายที่ตัดความหนาแน่นบริเวณที่ต่ำกว่าหรือเหนือกว่าศูนย์ออก ขึ้นอยู่กับเครื่องหมายที่กำหนด และสมมติให้สมาชิกของ $\Lambda$ ทุกตัว ทั้งที่ถูกกำหนดเครื่องหมายและไม่ได้ถูกกำหนดเครื่องหมายเป็นอิสระซึ่งกันและกัน มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับ $l_{oj}$ และค่าความแม่นยำ (ซึ่งเป็นส่วนกลับของความแปรปรวน) เท่ากับ $L_{oj}$ สมมติให้ $\gamma$ มีการแจกแจงแบบสม่ำเสมอไม่ตรงแบบ (Improper Uniform) และ $\Lambda$, $\phi$, $\Psi$ และ $\gamma$ เป็นอิสระซึ่งกันและกัน

ในการจัดอันดับโดยใช้การวิเคราะห์ปัจจัยเชิงเบส์ เราจะใช้ค่าคาดหวังภายใต้การแจกแจงความน่าจะเป็นภายหลังของคะแนนของปัจจัย ($\phi_i$) มาใช้ในการจัดอันดับ ซึ่งการประมาณค่าคาดหวังดังกล่าวสามารถทำได้จากการจำลองโดยใช้เทคนิคลูกโซ่มาร์คอฟ (MCMC) โดยในงานวิจัยนี้จะใช้การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampler) และการสุ่มตัวอย่างแบบฮิตแอนด์รัน (Hit-and-run Sampler) แล้วนำผลจากการจำลองมาเฉลี่ยเพื่อมาประมาณค่าคาดหวัง แต่เนื่องจากการประมาณค่าย่อมมีความผิดพลาด จึงใช้วิธีค่าเฉลี่ยกลุ่ม (batch mean method)

ประมาณค่าคลาดเคลื่อนมาตรฐาน (standard error) ของตัวประมาณ และนำค่าดังกล่าวมาใช้ในการเปรียบเทียบประสิทธิภาพของการสุ่มตัวอย่างแบบกิบส์ กับการสุ่มตัวอย่างแบบฮิตแอนด์รัน

### 2. ลูกโซ่มาร์คอฟ (Markov Chain)

พิจารณากระบวนการเฟ้นสุ่มแบบไม่ต่อเนื่อง (discrete-time stochastic process) $\{X_n; n = 0,1,2,\ldots\}$ กระบวนการนี้เป็นลูกโซ่มาร์คอฟแบบไม่ต่อเนื่อง (discrete-time Markov chain) ถ้า $m, n \geq 0$ และสถานะ $i, j \in S, x_u$ และ $0 \leq u < m$ เป็นจำนวนเต็มที่ไม่เป็นลบ (nonnegative integers)

$$P(X_{m+n} = j \mid X_m = i, X_u = x_u, 0 \leq u < m) = P(X_{m+n} = j \mid X_m = i)$$

ลูกโซ่มาร์คอฟแบบไม่ต่อเนื่อง หรือลูกโซ่มาร์คอฟ เป็นกระบวนการเฟ้นสุ่มที่มีคุณสมบัติมาร์คอฟ (Markov property) คือ ความน่าจะเป็นแบบมีเงื่อนไขของอนาคต $X_{m+n}$ เมื่อกำหนดปัจจุบัน $X_m$ และอดีต $X_u$ โดยที่ $0 \leq u < m$ จะขึ้นอยู่กับปัจจุบันเท่านั้น และเป็นอิสระกับอดีต $X_u$ เรียก $S$ ว่า สเปซของสถานะ (state space)

#### 2.1 ทฤษฎีลูกโซ่มาร์คอฟสำหรับสเปซทั่วไป

สำหรับลูกโซ่มาร์คอฟบนสเปซ $(S, \Im)$ และ $\{X_n; n = 0,1,2,\ldots\}$ เป็นลูกโซ่มาร์คอฟ ถ้า

$$P(X_{n+1} \in A \mid X_n, X_{n-1}, \ldots, X_0) = P(X_{n+1} \in A \mid X_n)$$

$$= P(X_n, A) \qquad \text{สำหรับ } A \in \Im$$

ซึ่ง $P(X_n, A)$ มีคุณสมบัติดังนี้

1. สำหรับแต่ละ $x \in \Im$ ฟังก์ชัน $P(x, \cdot)$ เมเซอร์ความน่าจะเป็น (measure probability) บน $(S, \Im)$
2. สำหรับแต่ละ $A \in \Im$ ฟังก์ชัน $P(\cdot, A)$ สามารถวัดได้ (measurable)

<u>นิยาม</u> ทฤษฎีบทจำกัด

$$\mu P(A) = \int_S P(x, A)\mu(dx)$$

เมื่อกำหนดความน่าจะเป็นในการเปลี่ยนสถานะขั้นที่ 1 คือ

$$P^1(x,A) = P(x,A)$$

และความน่าจะเป็นในการเปลี่ยนสถานะขั้นที่ $n$ คือ

$$P^n(x,A) = \int_S P^{n-1}(y,A)P(x,dy)$$

ดังนั้น

$$P(X_n \in A) = \mu P^n(A)$$

เมเซอร์ความน่าจะเป็น $\pi$ บน $(S,\Im)$ เป็นการแจกแจงคงตัว (stationary distribution) ของ $P$ ถ้า

$$\pi P = P$$

**2.2 ทฤษฎีบทการลู่เข้าของลูกโซ่มาร์คอฟในกรณี total variation (Markov Chain Convergence in Total Variation)**

ให้ $P$ เป็น $\varphi$ ที่ irreducible สำหรับบางค่า $\sigma -$ เมเซอร์ที่หาค่าได้ $\varphi$ บน $(S,\Im)$ ถ้า $P$ เป็น aperiodic แล้ว

$$\lim_{n\to\infty} P(X_n \in A \mid X_0 = x) = \pi(A) \text{ สำหรับทุก } A \in \Im$$

เมื่อ $\pi$ เป็นการแจกแจงคงตัว (stationary distribution) ของ $P$ ดังนั้น

$$\pi(\cdot) = \pi P(\cdot) = \int_S P(x,\cdot)\pi(dx)$$

1. The First Lyapunov Condition (FLC)

สำหรับเซตไม่ว่าง $B \subset S$ สเกลาร์ที่เป็นบวก (positive scalars) $a < 1$ $b$ และ $\delta$ เป็นจำนวนเต็ม $m \geq 1$ การแจกแจงความน่าจะเป็น $\varphi$ บน $S$ และฟังก์ชัน $V : S \to [1,\infty]$ จะได้ว่า

- $P(X_m \in \cdot \mid X_0 \in z) \geq \delta\varphi(\cdot)$ สำหรับทุก $z \in B$
- $E\big(V(X_1) \mid X_0 = z\big) \leq aV(z) + bI(z \in B)$ สำหรับทุก $z \in S$

หมายความว่า ถ้าลูกโซ่มาร์คอฟเป็นไปตามเงื่อนไข FLC และมีคุณสมบัติ aperiodic แล้วลูกโซ่มาร์คอฟนี้เป็น ergodic อย่างสม่ำเสมอ (Uniformly ergodic)

2. Markov Chain Strong Law of Large Numbers (MCSLLN)

ให้ $\{X_n; n = 0,1,2,\ldots\}$ เป็นลูกโซ่มาร์คอฟที่ ergodic อย่างสม่ำเสมอ (uniformly ergodic Markov chain) บนสเปซ $S$ ที่มี $\pi$ เป็นการแจกแจงคงตัว (stationary distribution) และฟังก์ชัน $h: S \to \Re$ ถ้า $\pi|h| = E_\pi[|h(X_0)|] = \int_S |h(x)|\pi(dx) < \infty$ แล้ว

$$\frac{1}{n}\sum_{i=0}^{n-1} h(X_i) \to \pi h = \int_S h(x)\pi(dx) \quad \text{เมื่อ } n \to \infty$$

3. Markov Chain Central Limit Theorem (MCCLT)

ให้ $\{X_n; n = 0,1,2,\ldots\}$ เป็นลูกโซ่มาร์คอฟที่ ergodic อย่างสม่ำเสมอ (uniformly ergodic Markov chain) บนสเปซ $S$ และฟังก์ชัน $h: S \to \Re$ ด้วยค่า $h^2(z) \le cV(z)$ สำหรับบางค่า $c > 0$ และทุกค่า $z$ จะได้ว่า

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n-1} h(X_i) - \pi h\right) \Rightarrow \sigma\, N(0,1)$$

เมื่อ $\pi$ เป็นการแจกแจงความน่าจะเป็นคงตัว (stationary probability distribution) ของ $X$ และ $\sigma^2 = Var_\pi[h(X_0)] + 2\sum_{k=1}^{\infty} \mathrm{cov}_\pi[h(X_0), h(X_k)]$

4. Reversibility

ถ้าลูกโซ่มาร์คอฟใดมีคุณสมบัติ reversibility แล้ว จะสามารถออกแบบเมทริกซ์ของความน่าจะเป็นในการเปลี่ยนสถานะ (transition probability matrix) ได้ง่ายขึ้น

<u>ทฤษฎีบท</u> ถ้าความน่าจะเป็นในการเปลี่ยนสถานะ $P$ มีคุณสมบัติ reversibility ตามความน่าจะเป็น $\pi$ แล้ว $\pi$ จะเป็นการแจกแจงคงตัวของ $P$

ให้ $A \subset S$ แล้ว

$$\begin{aligned}\pi P(A) &= \int_S P(x,A)\pi(dx) \\ &= \int_A P(x,S)\pi(dx) \\ &= \int_A 1\pi(dx) \\ &= \pi(A)\end{aligned}$$

### 2.3 การวิเคราะห์สถานะเสถียรภาพของการจำลอง (Steady State Analysis of a Simulation)

ในการจำลองกระบวนการเฟ้นสุ่มที่เป็นลูกโซ่มาร์คอฟ $\{X_n; n = 0,1,2,\ldots\}$ เมื่อ $X_n$ เป็นเวกเตอร์ของตัวแปรสุ่ม ซึ่ง $X_n \in S$ และ $S$ เป็นสเปซของสถานะ (state space) โดยเงื่อนไขทั่วไป การแจกแจงของ $X_n$ จะลู่เข้าสู่การแจกแจงคงตัว $\pi$ ซึ่งในการวิเคราะห์เชิงเบส์ก็คือการแจกแจงภายหลัง และเป็นอิสระกับจุดเริ่มต้น $x_0$ ดังนี้

$$p(X_n \in A \mid X_0 = x_0) \to \pi(A)$$

จาก MCSLLN จะได้ว่า ค่าเฉลี่ยในระยะยาว (long run average) จะสามารถประมาณค่า $\pi h$ ดังนี้

$$\frac{1}{N}\sum_{i=0}^{N-1} h(X_i) \to \pi h \text{ สำหรับ } N \text{ ที่มีค่ามาก}$$

จาก MCCLT จะได้ว่า การแจกแจงของค่าเฉลี่ยในระยะยาว (distribution of long run average) จะประมาณ (approximate) ด้วยการแจกแจงแบบปกติ (normal distribution) ดังนี้

$$\frac{1}{N}\sum_{i=0}^{N-1} h(X_i) \sim N(\pi h, \sigma^2) \text{ สำหรับ } N \text{ ที่มีค่ามาก}$$

เมื่อ $\sigma^2 = Var_\pi[h(X_0)] + 2\sum_{k=1}^{\infty} \mathrm{cov}_\pi[h(X_0), h(X_k)]$

### 3. เทคนิคลูกโซ่มาร์คอฟมอนติคาร์โล (Markov Chain Monte Carlo)

ในทฤษฎีบทของเบส์กรณีที่การแจกแจงภายหลัง(Posterior distribution) ไม่ได้อยู่ในรูปแบบสามัญนั้น การคำนวณค่าคาดหวังของพารามิเตอร์จากฟังก์ชันความหนาแน่นน่าจะเป็นภายหลังเป็นปัญหาสำคัญของการสร้างแบบจำลองเบส์ เทคนิคลูกโซ่มาร์คอฟมอนติคาร์โล (Markov Chain Monte Carlo) หรือ MCMC จึงได้รับการเสนอเพื่อแก้ปัญหาดังกล่าว โดย MCMC เป็นเทคนิคที่ใช้ในการสุ่มตัวอย่างของตัวแปรสุ่มจากฟังก์ชันความหนาแน่นน่าจะเป็นภายหลัง

MCMC เป็นการสร้างลำดับของเวคเตอร์สุ่ม $\{\vec{\theta}_0, \vec{\theta}_1, \ldots, \vec{\theta}_n\}$ โดยที่การสุ่ม $\vec{\theta}_{j+1}$ จากการแจกแจง $p(\vec{\theta}_{j+1} | \vec{\theta}_j)$ เมื่อ $j \geq 0$ ซึ่งหมายถึง $\vec{\theta}_{j+1}$ ถูกสุ่มโดยขึ้นอยู่กับ $\vec{\theta}_j$ เพียงอย่างเดียวไม่ได้ขึ้นอยู่กับ $\{\vec{\theta}_0, \vec{\theta}_1, \ldots, \vec{\theta}_{j-1}\}$ นั่นคือ $p(\vec{\theta}_{j+1} | \vec{\theta}_j)$ เป็น Transition Kernel ซึ่งเป็นความ

น่าจะเป็นของการเปลี่ยนสถานะจากสถานะที่ $j$ ไปยังสถานะที่ $j+1$ โดยเงื่อนไขทั่วไป เมื่อ $j$ มีค่ามากขึ้น การแจกแจงของ $\bar{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว $\pi$ (stationary distribution) ในกรณีการวิเคราะห์เชิงเบส์ $\pi$ ก็คือการแจกแจงภายหลัง

ในการสร้าง MCMC นั้นจะมีขั้นตอนวิธี (Algorithm) หลายแบบ เช่น การสุ่มตัวอย่างแบบเมโทรโพลิส-เฮสติงส์ (Metropolis-Hastings Sampling), การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampling) และ การสุ่มตัวอย่างแบบฮิตแอนด์รัน ซึ่งมีรายละเอียดดังต่อไปนี้

### 3.1 การสุ่มตัวอย่างแบบเมโทรโปลิส-เฮสติงส์ (Metropolis-Hastings Sampling)

สมมติว่าต้องการสุ่มพารามิเตอร์ $\bar{\theta}$ จากการแจกแจง $\pi$ ซึ่งก็คือการแจกแจงภายหลัง (posterior distribution) ของการวิเคราะห์เชิงเบส์ ในการกระบวนการ (Algorithm) ของเมโทรโปลิส-เฮสติงส์มีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 กำหนดพารามิเตอร์เริ่มต้น $\bar{\theta}_0$ โดยที่ $\pi > 0$

ขั้นที่ 2 สุ่มพารามิเตอร์ $\bar{\theta}^{(can)}$ จาก Proposal Distribution $q\left(\bar{\theta}_j \middle| \bar{\theta}_{j+1}\right)$ ซึ่งเป็นความน่าจะเป็นในการเปลี่ยนค่าพารามิเตอร์จาก $\bar{\theta}_j$ ไปเป็น $\bar{\theta}_{j+1}$

ขั้นที่ 3 คำนวณหาค่า $\alpha$ ซึ่งเป็นอัตราส่วนของฟังก์ชันความหนาแน่น $\pi$ ของพารามิเตอร์ที่สุ่มขึ้นมาใหม่ $\bar{\theta}^{(can)}$ เทียบกับฟังก์ชันความหนาแน่นของพารามิเตอร์ $\bar{\theta}_j$

$$\alpha = \min\left( \frac{\pi\left(\bar{\theta}^{(can)}\right) q\left(\bar{\theta}^{(can)} \middle| \bar{\theta}_j\right)}{\pi\left(\bar{\theta}_j\right) q\left(\bar{\theta}_j \middle| \bar{\theta}^{(can)}\right)},\ 1 \right)$$

ขั้นที่ 4 สุ่ม $u$ จากการแจกแจงยูนิฟอร์ม (Uniform Distribution) $U(0,1)$

ยอมรับ $\bar{\theta}^{(can)}$ ถ้า $u < \alpha$ และให้ $\bar{\theta}_{j+1} = \bar{\theta}^{(can)}$

หรือ ปฏิเสธ $\bar{\theta}^{(can)}$ ถ้า $u \geq \alpha$ และให้ $\bar{\theta}_{j+1} = \bar{\theta}_j$

ขั้นที่ 5 วนซ้ำในขั้นที่ 2 ใหม่

โดยเงื่อนไขทั่วไป การแจกแจงของ $\bar{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว $\pi$ (stationary distribution) ซึ่งสามารถนำไปประมาณค่าคาดหวังภายใต้การแจกแจงภายหลังได้

### 3.2 การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampling)

การสุ่มตัวอย่างแบบกิบส์เป็นกรณีพิเศษของ Single-component Metropolis Hasting Method กล่าวคือจะสุ่ม $\vec{\theta}^{(can)}$ จาก Full Condition Distribution แทนการสุ่มจาก Proposal Distribution นั่นคือ

$$q\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_j,\vec{\theta}_{-j}\right)=\pi\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)$$

และ $$q\left(\vec{\theta}_j\middle|\vec{\theta}_j^{(can)},\vec{\theta}_{-j}\right)=\pi\left(\vec{\theta}_j\middle|\vec{\theta}_{-j}\right)$$

ซึ่งทำให้

$$\alpha=\min\left(\frac{\pi\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)q\left(\vec{\theta}_j\middle|\vec{\theta}_j^{(can)},\vec{\theta}_{-j}\right)}{\pi\left(\vec{\theta}_j|,\vec{\theta}_{-j}\right)q\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_j,\vec{\theta}_{-j}\right)},\ 1\right)$$

$$=\min\left(\frac{\pi\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)p\left(\vec{\theta}_j\middle|\vec{\theta}_{-j}\right)}{\pi\left(\vec{\theta}_j|,\vec{\theta}_{-j}\right)p\left(\vec{\theta}_j^{(can)}\middle|\vec{\theta}_{-j}\right)},\ 1\right)=1$$

การสุ่ม $\vec{\theta}_j^{(can)}$ จาก Full Conditional Distribution จึงถูกยอมรับในทุกๆครั้ง

ดังนั้น การสุ่มพารามิเตอร์ในการสุ่มตัวอย่างแบบกิบส์จึงเป็นการสุ่มจาก Full Conditional Distribution เป็นดังนี้

$$\pi\left(\vec{\theta}_j\middle|\vec{\theta}_{-j}\right)=\frac{\pi\left(\vec{\theta}\right)}{\int\pi\left(\vec{\theta}\right)d\theta_j}\text{ โดยที่ }\vec{\theta}_{-j}=\left\{\theta_1,\theta_2,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_J\right\}$$

โดยเงื่อนไขทั่วไป การแจกแจงของ $\vec{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว $\pi$ (stationary distribution) ซึ่งสามารถนำไปประมาณค่าคาดหวังภายใต้การแจกแจงภายหลังได้

### 3.3 การสุ่มตัวอย่างแบบฮิตแอนด์รัน (Hit-and-run sampler)

ให้ $S\subset\Re^n$ และให้การแจกแจงเป้าหมาย (Target distribution) มีฟังก์ชันความน่าจะเป็น $\pi$ บน $S$

สร้าง $\vec{\theta}_n$ ด้วยกระบวนการดังนี้

ขั้นที่ 1 กำหนดพารามิเตอร์เริ่มต้น $\vec{\theta}_0\in S$

ขั้นที่ 2 ที่ $\vec{\theta}_n=\vec{\theta}\in S$ สุ่มเลือกทิศทาง $\vec{d}$ บนพื้นผิวทรงกลมรัศมี 1 หน่วย ลากเส้นตรง $L$ ผ่าน $\vec{\theta}$ และตัดกับ $S$ โดยให้ $L_n=S\cap\left\{l:l=\vec{\theta}+r\vec{d},r\in\Re\right\}$

ขั้นที่ 3 สุ่มเลือกจุด $\bar{\theta}_{n+1}$ บนเส้นตรง $L_n$ ด้วยการแจกแจงแบบ $\pi$ โดยมีเงื่อนไขบน $L_n$

ขั้นที่ 4 วนซ้ำในขั้นที่ 2 ใหม่

โดยเงื่อนไขทั่วไป การแจกแจงของ $\bar{\theta}_j$ จะลู่เข้าสู่การแจกแจงคงตัว $\pi$ (stationary distribution) ซึ่งสามารถนำไปประมาณค่าคาดหวังภายใต้การแจกแจงภายหลังได้

จากการลู่เข้าของลูกโซ่มาร์คอฟจึงมีค่าความแปรปรวนร่วม (covariance) เพราะ $\bar{\theta}_j$ แต่ละตัวไม่เป็นอิสระกัน ดังนั้น งานวิจัยนี้จึงใช้วิธีค่าเฉลี่ยกลุ่ม (Batch Means method) สำหรับแก้ปัญหาค่าความแปรปรวนร่วม

**4. วิธีค่าเฉลี่ยกลุ่ม** (Batch means method)

เป็นวิธีการหนึ่งที่ใช้ประมาณค่าเฉลี่ยประชากร โดยใช้การคำนวณ Monte Carlo standard error (MCSE) สมมติว่าเราสนใจประมาณค่าเฉลี่ย $E_\pi\left(g\left(\bar{\theta}\right)\right)$ เมื่อ $\bar{\theta}$ มีการแจกแจง $\pi$ ซึ่งกระบวนการของค่าเฉลี่ยกลุ่ม (Batch means) มีดังนี้

ขั้นที่ 1 จำลองลูกโซ่มาร์คอฟ $\left\{\bar{\theta}_n\right\}$ จำนวน $n = ab$ รอบ โดยที่ $a$ และ $b$ เป็นจำนวนเต็ม $a$ เป็นจำนวน batch และแต่ละ batch มีขนาด $b$

ขั้นที่ 2 หาค่าเฉลี่ยของตัวอย่าง ใน batch ที่ $k$ จากสูตร $\bar{Y}_k = \frac{\sum_{i=(k-1)b+1}^{kb} g\left(\bar{\theta}_i\right)}{b}$

ขั้นที่ 3 กำหนดให้ $\hat{\sigma}_g^2 = \frac{b}{a-1}\sum_{k=1}^{a}\left(\bar{Y}_k - \bar{g}_n\right)^2$ โดยที่ $\bar{g}_n = \frac{1}{a}\sum_{k=1}^{a}\bar{Y}_k$ จะได้ว่าค่าประมาณของ Monte Carlo standard error คือ $\frac{\hat{\sigma}_g}{\sqrt{n}}$

ขั้นที่ 4 คำนวณหาช่วงความเชื่อมั่นของ $E_\pi\left(g\left(\bar{\theta}\right)\right)$ ได้จากสูตร

$$\bar{g}_n \pm t_{\frac{\alpha}{2},(a-1)}\frac{\hat{\sigma}_g}{\sqrt{n}}$$



**เอกสารและงานวิจัยที่เกี่ยวข้อง**

Chen และ Schmeiser (1993) ได้ทำการเปรียบเทียบประสิทธิภาพการสุ่มตัวอย่างแบบฮิตแอนด์รัน และการสุ่มตัวอย่างแบบกิบส์ พบว่าการสุ่มตัวอย่างแบบฮิตแอนด์รันมีประสิทธิภาพมากว่าสำหรับในกรณีที่เป็นการแจกแจงแบบปกติของ 2 ตัวแปร

Quinn (2004) ได้นำเสนอการใช้ตัวแบบการวิเคราะห์ปัจจัยเชิงเบส์ (Bayesian factor analysis) ซึ่งได้ปรับให้เหมาะกับการวิเคราะห์ข้อมูลแบบผสมที่มีทั้งตัวแปรแบบต่อเนื่องและตัวแปรเชิงอันดับ และใช้วิธีการคำนวณที่วางอยู่บนพื้นฐานของการสุ่มตัวอย่างแบบกิบส์ และวิธีของ Cowles (1993) ซึ่งวิธีดังกล่าวจัดอยู่ในกลุ่มวิธีลูกโซ่มาร์คอฟ มอนติคาร์โล (Markov chain Monte Carlo) หรือ MCMC ในการประมาณค่าพารามิเตอร์

Lovasz และ Vempala (2006) ได้ศึกษาการสุ่มตัวอย่างแบบฮิตแอนด์รันบนบริเวณคอนเวกซ์ พบว่า การสุ่มตัวอย่างแบบฮิตแอนด์รันเป็นหนึ่งในวิธีที่เร็วที่สุด ในการสร้างจุดตัวอย่างซึ่งลู่เข้าสู่การแจกแจงแบบสม่ำเสมอบน Log-concave

จึงเป็นที่น่าสนใจว่า หากแทนคำนวณวิธีการสุ่มตัวอย่างแบบกิบส์ด้วยการสุ่มตัวอย่างแบบฮิตแอนด์รันในตัวแบบการวิเคราะห์ปัจจัยเชิงเบส์ที่ Quinn ได้นำเสนอ ผลที่ได้ยังคงสอดคล้องกับงานวิจัยของ Chen และ Schmeiser และงานวิจัยของ Lovasz และ Vempala หรือไม่