5 吉氏取樣器(Gibbs sampler)的介紹

5.1 吉氏取樣器

吉氏取樣器是一種利用條件分配間接得到邊際分配的工具，它避免掉了在算邊際分配時所遇到的複雜的積分問題，取而代之的是一個以較簡單的計算就可以得到的數列，以下就針對吉氏取樣器做簡單的介紹。

假設我們給定一個聯合密度函數 $f(x_1, x_2, \ldots, x_p)$，且已知條件機率密度函數 $f(x_s \mid x_1, x_2, \ldots, x_{s-1}, x_{s+1}, \ldots, x_p)$。如果我們想要得到它的邊際分配，最一般的作法即是利用下列的式子

$$f(x_s) = \int \ldots \int f(x_1, x_2, \ldots, x_p) dx_1 dx_2 \ldots dx_{s-1} d_{s+1} \ldots dx_p \text{，} \tag{5.1}$$

將邊際分配 $f(x_s)$ 算出。但實際上，在許多的情況下，(5.1) 式是很難積分出來的。此時，我們便可以利用吉氏取樣器來幫助我們間接的得到邊際分配 $f(x_s)$。

吉氏取樣器的過程如下

1. 先給定任意的初始值 $x_1^{(0)}, x_2^{(0)}, \ldots, x_{p-1}^{(0)}$，
   從 $f(x_p \mid X_1 = x_1^{(0)}, \ldots, X_{p-1} = x_{p-1}^{(0)})$ 分配中隨機取出 $x_p^{(0)}$。

2. 從 $f(x_1 \mid X_2 = x_2^{(0)}, \ldots, X_p = x_p^{(0)})$ 分配中隨機取出 $x_1^{(1)}$，
   再從 $f(x_2 \mid X_1 = x_1^{(1)}, X_3 = x_3^{(0)}, \ldots, X_p = x_p^{(0)})$ 分配中隨機取出 $x_2^{(1)}$，
   以此類推，直到從 $f(x_p \mid X_1 = x_1^{(1)}, \ldots, X_{p-1} = x_{p-1}^{(1)})$ 分配中隨機取出 $x_p^{(1)}$，此時我們稱完成一次疊代。

3. 在 k 次疊代之後，我們可以得到 $(x_1^{(k)}, x_2^{(k)}, \ldots, x_p^{(k)})$。

Geman and Geman(1984) 指出，在適當的條件下，

$$X_s^{(k)} \to X_s, \text{ as } k \to \infty \text{，}$$

其中X_s服從$f(x_s)$的分配。所以當k夠大時，我們將$x_s^{(k)}$視為X_s的模擬觀察值。

吉氏取樣器讓我們避免掉在積分計算上的困難，只需要利用簡單的計算生成一組序列，經過足夠次數的疊代後，就能得到模擬邊際分配的觀察值。在利用電腦模擬出大量的樣本後，我們便可以利用這樣一組樣本來估計邊際分配的期望值或是變異數等等與邊際相關的資訊。

以兩個隨機變數(X, Y)為例，吉氏取樣器經由交替地從條件分配$f(x \mid y)$和$f(y \mid x)$產生樣本，可以生成一個X的樣本，其中X服從$f(x)$的分配。過程中形成一個由隨機變數所組成的"Gibbs sequence" 如下

$$Y^{(0)}, X^{(0)}, Y^{(1)}, X^{(1)}, Y^{(2)}, X^{(2)}, \ldots, Y^{(k)}, X^{(k)}\text{，} \tag{5.2}$$

其中除了$Y^{(0)}$是隨意給定的初始值外，其餘皆以下列規則產生

$$\begin{aligned} X^{(j)} &\sim f(x \mid Y = y^{(j)})\text{，} \\ Y^{(j+1)} &\sim f(y \mid X = x^{(x)})\text{。} \end{aligned} \tag{5.3}$$

且當$k \to \infty$，則$X^{(k)}$的分配會收斂到X的真正的邊際分配$f(x)$。

5.2 簡單的收斂說明

設X, Y分別為Bernoulli隨機分配，其聯合機率密度函數如下

$$\begin{bmatrix} f_{xy}(0,0) & f_{xy}(1,0) \\ f_{xy}(0,1) & f_{xy}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}\text{。}$$

對於此分配，我們可以看出x的邊際分配

$$f_x = [f_x(0), f_x(1)] = [p_1 + P_3, p_2 + p_4]\text{，}$$

是一個成功機率為$p_2 + p_4$的Bernoulli分配。

則我們可以很直接就得到$X \mid Y = y$和$Y \mid X = x$兩個條件機率密度函數

$$\boldsymbol{A}_{y|x} = \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix} = \begin{bmatrix} P(Y=0 \mid X=0) & P(Y=1 \mid X=0) \\ P(Y=0 \mid X=1) & P(Y=1 \mid X=1) \end{bmatrix},$$

及

$$\boldsymbol{A}_{x|y} = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix} = \begin{bmatrix} P(X=0 \mid Y=0) & P(X=1 \mid Y=0) \\ P(X=0 \mid Y=1) & P(X=1 \mid Y=1) \end{bmatrix}。$$

前面矩陣中$P(Y = y \mid X = x)$表示由x轉換到y的機率，所以我們可以將矩陣$\boldsymbol{A}_{y|x}$視爲一轉換矩陣，爲給定由x的狀況轉換至y的狀況的機率。同理，$\boldsymbol{A}_{x|y}$爲一由y轉換到x的轉換矩陣。

若我們只想知道X的邊際分配，則在(5.2)的數列中，我們可以很清楚的看出，從$X^{(0)} \to X^{(1)}$的過程中，必須經過$Y^{(1)}$，這樣的過程形成了一個馬可夫鏈，轉換機率爲

$$P(X = x^{(1)} \mid X = x^{(0)}) = \sum_y P(X = x^{(x)} \mid Y = y) \times P(Y = y \mid X = x^{(0)})。$$

此時X數列的轉換矩陣可表示爲

$$\boldsymbol{A}_{x|x} = \boldsymbol{A}_{x|y}\boldsymbol{A}_{y|x}。$$

因此我們便能簡單的計算出數列中$X^{(k)}$的機率分配。也就是說，$P(X = x^{(k)} \mid X = x^{(0)})$的轉換矩陣爲$(\boldsymbol{A}_{x|x})^k$。若我們以$\boldsymbol{f}_k = [f_k(0) \quad f_k(1)]$來表示$X^{(k)}$的邊際機率密度函數，則對於所有$k$，

$$\boldsymbol{f}_k = \boldsymbol{f}_0 \boldsymbol{A}_{x|x}^{k-1} \boldsymbol{A}_{x|x} = \boldsymbol{f}_{k-1}\boldsymbol{A}_{x|x}$$

Hoel, Port, and Stone(1972)提到，只要$\boldsymbol{A}$的元素全爲正數，則當$k \to \infty$，$\boldsymbol{f}_k$會收斂到唯一的分配$\boldsymbol{f}$，並滿足

$$\boldsymbol{f}\boldsymbol{A}_{x|x} = \boldsymbol{f} \tag{5.4}$$

所以，若Gibbs sequence收斂，則滿足(5.4)式中的$\boldsymbol{f}$必爲X的邊際分配。

當然，在實際操作上，我們無法將k取到無窮大，那k要取到多少才適當，目前尚未有一肯定答案。在此，我們採用Gelfand and Smith(1990)提出的方法，獨立生成m個長度爲k的Gibbs sequences，然後取每個數列最後一個$X^{(k)}$。只要k夠大，則將$X^{(k)}$視爲X獨立的模擬樣本，其中X服從$\boldsymbol{f}(x)$的分配。Gelfand and Smith(1990)所處理的$k \leq 50$，$m \leq 1000$，而且他們認爲k不需要大於50。

5.3 吉氏取樣器在不完整多元伯努利上的應用

現在將吉氏取樣器應用於失去部分訊息資料(censored data)上。

假設我們收到n筆資料，令爲$\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$。$\boldsymbol{\theta} \sim D(\boldsymbol{a})$及$\lambda_{i*} \sim D(b_{i*})$，其中$i = 1, 2, \ldots, I$爲先驗分配。$\boldsymbol{w} = \boldsymbol{\theta} \cdot \boldsymbol{\Lambda}$，其中第$j$個元素$w_j$表示報告爲$j$的機率，$j = 1, 2, \ldots, J$。再假設這$n$筆報告的受訪者真實類別爲$\boldsymbol{Y}$，$\boldsymbol{Y}$是一個$1 \times n$的隨機向量，其中$Y_k$表示爲第$k$位受訪者的真實類別。在此，我們須先給定$\boldsymbol{Y}$的起始值，表示爲$\boldsymbol{y}^{(0)} = (y_1^{(0)}, y_2^{(0)}, \ldots, y_n^{(0)})$。我們所取的起始值的方法是假設對第$k$個觀察值而言，在收到$R_k = r_k$的情況下，其真實類別$Y_k$會服從多元伯努利分配，其中各類別機率向量如下

$$\left(\frac{\theta_1^{(0)} \cdot \lambda_{1r_k}^{(0)}}{\sum_{m=1}^{I} \theta_m^{(0)} \cdot \lambda_{mr_k}^{(0)}}, \frac{\theta_2^{(0)} \cdot \lambda_{2r_k}^{(0)}}{\sum_{m=1}^{I} \theta_m^{(0)} \cdot \lambda_{mr_k}^{(0)}}, \ldots, \frac{\theta_I^{(0)} \cdot \lambda_{Ir_k}^{(0)}}{\sum_{m=1}^{I} \theta_m^{(0)} \cdot \lambda_{mr_k}^{(0)}} \right) 。$$

利用此機率分配，我們可以生成$y_k^{(0)}$，$k = 1, 2, \ldots, n$。

接著我們依照下列步驟進行：

步驟一、首先對$\boldsymbol{y}^{(0)}$，$\boldsymbol{r}$作一個統計，定義下列矩陣

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_I) \text{，} \boldsymbol{z} = \begin{bmatrix} z_{11} & z_{12} & \ldots & z_{1J} \\ z_{21} & z_{22} & \ldots & z_{2J} \\ \vdots & \vdots & & \vdots \\ z_{I1} & z_{I2} & \ldots & z_{IJ} \end{bmatrix} ,$$

其中x_i表示$\boldsymbol{y}=(y_1,y_2,\ldots,y_n)$中屬於類別$i$的數量。$z_{ij}$表示$n$個配對$[y_k,r_k]$(即[真實類別，報告值])中$y_k=i,r_k=j$的數量，$k=1,2,\ldots,n$。因此第一個步驟後我們可以得到$\boldsymbol{x}^{(0)}$，$\boldsymbol{z}^{(0)}$。

步驟二、利用$\boldsymbol{x}^{(0)}$，$\boldsymbol{z}^{(0)}$來計算下列兩數：

$$\begin{aligned}\boldsymbol{a}^{(1)} &= \boldsymbol{a}^{(0)}+\boldsymbol{x}^{(0)}=(a_1^{(0)}+x_1^{(0)},\ldots,a_I^{(0)}+x_I^{(0)})=(a_1^{(1)},a_2^{(1)},\ldots,a_I^{(1)})\text{，}\\ \boldsymbol{b}^{(1)} &= \boldsymbol{b}^{(0)}+\boldsymbol{z}^{(0)}=[(b_{ij}^{(0)}+z_{ij}^{(0)})]_{I\times J}=[(b_{ij}^{(1)})]_{I\times J}\text{。}\end{aligned}$$

步驟三、利用Jiang, Kadane, and Dickey(1992)所提出的Monte Carlo方法來模擬$\boldsymbol{\theta}$以及$\boldsymbol{\lambda}$的觀察值，方法如下：

1. 利用自由度爲$2a_i^{(1)}$的卡方分配來生成I個隨機變數$u_i^{(1)}$，$i=1,2,\ldots,I$。
2. 令$\theta_i^{(1)}=\frac{u_i^{(1)}}{\sum_{m=1}^{I}u_m^{(1)}}$，此時可將$\boldsymbol{\theta}^{(1)}$視爲參數爲$\boldsymbol{a}^{(1)}$之Dirichlet分配的觀察值。

同樣的，對每個i，$i=1,2,\ldots,I$，當i固定的情況下，

1. 利用自由度爲$2b_{ij}^{(1)}$的卡方分配來生成J個隨機變數$g_{ij}^{(1)}$，$j=1,2,\ldots,J$。
2. 令$\lambda_{ij}^{(1)}=\frac{g_{ij}^{(1)}}{\sum_{m=1}^{J}g_{im}^{(1)}}$，此時可將$\boldsymbol{\lambda}_{i*}^{(1)}$視爲參數爲$\boldsymbol{b}_{i*}^{(1)}$之Dirichlet分配的觀察值。

步驟四、此一步驟又分成n個小步驟。對第一個觀察值而言，在收到$R_1=r_1$的情況下，真實類別$\boldsymbol{Y}_1$會服從多元伯努利分配，其中各類別機率向量如下：

$$\left(\frac{\theta_1^{(1)}\cdot\lambda_{1r_1}^{(1)}}{\sum_{m=1}^{I}\theta_m^{(1)}\cdot\lambda_{mr_1}^{(1)}},\frac{\theta_2^{(1)}\cdot\lambda_{2r_1}^{(1)}}{\sum_{m=1}^{I}\theta_m^{(1)}\cdot\lambda_{mr_1}^{(1)}},\ldots,\frac{\theta_I^{(1)}\cdot\lambda_{Ir_1}^{(1)}}{\sum_{m=1}^{I}\theta_m^{(1)}\cdot\lambda_{mr_1}^{(1)}}\right)\text{。}$$

利用此機率分配，我們可生成$y_1^{(1)}$。同樣的，對第2個觀察值而言，在收到$R_2=r_2$的情況下，真實類別$\boldsymbol{Y}_2$會服從多元伯努利分配，其中各類別機率向量如下：

$$\left(\frac{\theta_1^{(1)}\cdot\lambda_{1r_2}^{(1)}}{\sum_{m=1}^{I}\theta_m^{(1)}\cdot\lambda_{mr_2}^{(1)}},\frac{\theta_2^{(1)}\cdot\lambda_{2r_2}^{(1)}}{\sum_{m=1}^{I}\theta_m^{(1)}\cdot\lambda_{mr_2}^{(1)}},\ldots,\frac{\theta_I^{(1)}\cdot\lambda_{Ir_2}^{(1)}}{\sum_{m=1}^{I}\theta_m^{(1)}\cdot\lambda_{mr_2}^{(1)}}\right)\text{。}$$

利用此機率分配，我們再生成$y_2^{(1)}$。以此類推，生成$y_k^{(1)}$，$k=1,2,\ldots,n$之後，得到$\boldsymbol{y}^{(1)}=(y_1^{(1)},y_2^{(1)},\ldots,y_n^{(1)})$。如此便完成一次疊代。

接下來進行第二次疊代，重複上列步驟，即

1. 利用$\boldsymbol{y}^{(1)}$計算出$\boldsymbol{x}^{(1)}$、$\boldsymbol{r}^{(1)}$。
2. 把$\boldsymbol{x}^{(1)}$、$\boldsymbol{r}^{(1)}$分別與$\boldsymbol{a}^{(0)}$、$\boldsymbol{b}^{(0)}$相加形成$\boldsymbol{a}^{(2)}$、$\boldsymbol{b}^{(2)}$。
3. 將$2a_i^{(2)}$、$2b_{ij}^{(2)}$當成自由度以Monte Carlo生成$\boldsymbol{\theta}^{(2)}$、$\boldsymbol{\lambda}_{i*}^{(2)}$。
4. 用$\boldsymbol{\theta}^{(2)}$、$\boldsymbol{\lambda}_{i*}^{(2)}$計算出第$k$個報告其多元伯努利分配中各類別的機率，再生成$y_k^{(2)}$，$k=1,2,\ldots,n$，完成第二次疊代。

依照同樣步驟繼續疊代下去，經過k次疊代後，若將$\boldsymbol{\theta}$，$\boldsymbol{\lambda}$視為(5.2)中的X，真實類別Y視為(5.2)中的Y，則可形成數列如下：

$$\boldsymbol{y}^{(0)}\rightarrow\boldsymbol{\theta}^{(1)},\boldsymbol{\lambda}^{(1)}\rightarrow\boldsymbol{y}^{(1)}\rightarrow\boldsymbol{\theta}^{(2)},\boldsymbol{\lambda}^{(2)}\rightarrow\cdots\rightarrow\boldsymbol{\theta}^{(k)},\boldsymbol{\lambda}^{(k)}\text{，}$$

當$k\rightarrow\infty$時，$\boldsymbol{\theta}^{(k)},\boldsymbol{\lambda}^{(k)}$會趨近於真實的機率。

依照Gelfand and Smith(1990)所提出的方法，我們採用疊代k次，取最後一次疊代的值$\boldsymbol{\theta}^{(k)}$，$\boldsymbol{\lambda}^{(k)}$。然後重新給定一起始值，再疊代$k$次，同樣也取最後一次的估計值。我們總共給$m$次起始值，所以也會有$m$個不同的$\boldsymbol{\theta}^{(k)}$，$\boldsymbol{\lambda}^{(k)}$，將這些數的平均做為Gibbs對$\boldsymbol{\theta}$，$\boldsymbol{\lambda}$的估計值。例如：

$$E(\theta)\doteq\frac{\sum_{i=1}^{m}\theta_i}{m}\text{，}$$

$$E(\theta^2)\doteq\frac{\sum_{i=1}^{m}\theta_i^2}{m}\text{。}$$

依照柯力文(2003)對選取k值的大小所做的模擬結果，發現k值取到20時差異已經不大，所以在本文我們沿用柯力文(2003)對k與m的選取，取$k=20$，$m=500$來做模擬。