Bayesian Inference with ecological applications Chapter 10

潜在的な事象を扱うための多項分布モデル *Latent Multinomial Models*

本章では、記録した頻度データが多項分布に従う潜在的な確率変数を集約したものと考えられるときの、データ解析の手法を紹介する。この手法は標識再捕獲の解析に幅広く応用でき、観察された捕獲履歴が複数の事象からなる場合には常に適切な手法である。個体の誤識別に関するモデルについて考えよう。

10.1 MODEL M_t

ある動物の閉鎖個体群における個体数 N の推定を行うことを考える。Model M_t は、その動物の捕獲確率 P_t はどの調査回 t でも個体や時間によって変化しないことを想定したモデルである。

個々の動物の捕獲履歴 capture history $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_T)$ は、総調査回数 T の長さをもつ二値のベクトルにより表わされる。例えば 4 回の調査で 1 回目と 3 回目のみに発見できた場合 {1,0,1,0} もしくは 1010 と記述される。捕獲履歴は

$$j \equiv j(\boldsymbol{\omega}) = 1 + \sum_{t=1}^{T} \omega_t \, 2^{t-1}$$

を用いて指標化すると扱いやすい。上記の $\boldsymbol{\omega} = 1010$ の例では $j = 6$ となる。

捕獲履歴 j をもつ個体の数を f_i とするとき、$\boldsymbol{f} = (f_1, f_2, \cdots, f_{2^T})$ は個体数 N とセル確率

$$\pi_j = \prod_{t=1}^{T} {p_t}^{\omega_{j,t}} \, (1 - p_t)^{1-\omega_{j,t}}$$

$\left(\omega_{j,t} = \text{捕獲履歴 } j \text{ における調査回 } t \text{ での記録} \right)$

を母数とする多項分布に従う確率変数である。

Model M_t による解析は容易である。観察履歴頻度 observed frequency $\boldsymbol{f}^+ = (f_2, f_3, \cdots, f_{2^T})$ は母数に観察個体数 n とセル確率

$${\pi_j}^+ = \frac{\pi_j}{1 - \pi_1}$$

をもつ多項分布に従っている。また観察個体数 n は二項分布 $n \sim \mathrm{B}(N, 1 - \pi_1)$ に従う。多項分布に従う $\boldsymbol{f}$ はそれらの積 $[\boldsymbol{f}] \propto [\boldsymbol{f}^+ | n]\,[n]$ として表わされるだろう。PANEL 10.1 に BUGS code を示す。

10.2 MODEL $\mathrm{M}_{t,\alpha}$

個体が正しく識別できない場合、データに存在しないのに観察される個体 ghost record が含まれるため、Model M_t は個体数を過大評価してしまう。そこで Model $\mathrm{M}_{t,\alpha}$ では、誤識別された個体は 1 個体しか存在しないと想定*して、誤識別の可能性をモデルに考慮する。正識別率 correct identification probability を α とすると、Model $\mathrm{M}_{t,\alpha}$ では以下のように 3 つの事象とその確率が考えられる。

$\tilde{\omega}_t = 0$	捕獲されない	確率 $1 - p_t$
$\tilde{\omega}_t = 1$	捕獲され、識別も正しい	確率 $\alpha\, p_t$
$\tilde{\omega}_t = 2$	捕獲されたが、識別が誤っている	確率 $(1 - \alpha)\, p_t$

* 遺伝物質を用いた識別では妥当なようです。

したがって、このモデルでは 3^T の長さをもつ潜在履歴 latent history $\{\tilde{\omega}_1, \tilde{\omega}_2, \cdots, \tilde{\omega}_T\}$ が存在し、同様に

$$i = 1 + \sum_{t=1}^{T} \tilde{\omega}_t \, 3^{t-1}$$

と指標化して扱うことができる。以下では、潜在的に起こりうる履歴を潜在履歴 latent history i 、実際にデータとして記録された履歴を記録履歴 recorded history j ということにする。

Model $\mathrm{M}_{t,\alpha}$ の解析では、始めに潜在履歴頻度 latent history frequency $\boldsymbol{X} = (x_1, x_2, \cdots, x_{3^T})'$ を記録履歴頻度 $\boldsymbol{f}^+ = (f_2, f_3, \cdots, f_{2^T})$ に変換する必要がある。

$$\boldsymbol{f}^+ = \mathbf{A}' \mathbf{X}$$

行列 $\mathbf{A}$ は、潜在履歴 i から記録履歴 j が生じるとき $\mathrm{A_{ij}} = 1$ 、それ以外で $\mathrm{A_{ij}} = 0$ となる $3^T \times (2^T - 1)$ の行列である（Table 10.1）。ここで、潜在履歴頻度の妥当な集合 $\{\boldsymbol{X}: \boldsymbol{A}'\boldsymbol{X} = \boldsymbol{f}^+\}$ を知ることができるのであれば、尤度は次式の条件付き確率の和として表現できる。

$$L(p_1, p_2, \cdots, p_T, \alpha \mid \boldsymbol{f}^+) = \sum_{\boldsymbol{X}:\boldsymbol{A}'\boldsymbol{X}=\boldsymbol{f}^+} \Pr(\boldsymbol{X} \mid p_1, p_2, \cdots, p_T, \alpha)$$

集合 $\{\boldsymbol{X}: \boldsymbol{A}'\boldsymbol{X} = \boldsymbol{f}^+\}$ の推定には線形代数の知識を必要とする。

Digression: Some Linear Algebra

$\mathbf{0}$ でない行列 $\boldsymbol{B}$ を $\mathbf{0}$ に写像するベクトルの集合、零空間 null space という概念を導入しよう。ベクトル $\boldsymbol{v}$ が行列 $\boldsymbol{B}$ の零空間に属する $\boldsymbol{B}\boldsymbol{v} = \mathbf{0}, \boldsymbol{v} \in \mathrm{null}(\boldsymbol{B})$ とき、$\boldsymbol{B}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y}$ から次のように記述できる。

$$\boldsymbol{B}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y} \;\Leftrightarrow\; \boldsymbol{B}(\boldsymbol{x} - \boldsymbol{y}) = 0 \;\Leftrightarrow\; \boldsymbol{x} = \boldsymbol{y} + \boldsymbol{v}$$

同様に、もし $\boldsymbol{A}'\boldsymbol{X}_\mathbf{0} = \boldsymbol{f}^+$ を満たす潜在履歴頻度のひとつ $\boldsymbol{X}_\mathbf{0}$ を見つけることができるなら

$$\{\boldsymbol{X}: \boldsymbol{A}'\boldsymbol{X} = \boldsymbol{f}^+\} \;\Leftrightarrow\; \{\boldsymbol{X}: \boldsymbol{A}'\boldsymbol{X} = \boldsymbol{A}'\boldsymbol{X}_\mathbf{0}\} \;\Leftrightarrow\; \{\boldsymbol{X}: \boldsymbol{X} = \boldsymbol{X}_\mathbf{0} + \boldsymbol{v}, \boldsymbol{v} \in \mathrm{null}(\boldsymbol{A}')\}$$

と記述できる。$\boldsymbol{X}_\mathbf{0}$ となる潜在履歴頻度は、個体の識別が完全に正しい場合($\alpha = 1$) に容易に得られる。

行列 $\boldsymbol{A}'$ の零空間は、線形に独立な $r = 3^T - 2^T + 1$ 個の基底ベクトル basis vector $\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_r$ からなり、$a_1, a_2, \cdots, a_r$ を定数とすると

$$\boldsymbol{v} = \sum_{k=1}^{r} a_k \boldsymbol{b}_k \;, \forall\, \boldsymbol{v} \in \mathrm{null}(\boldsymbol{B})$$

が成り立つ（Table 10.2）。

Sample Calculation of Null Space Basis Vector

$T = 2$ の場合について、行列 $\boldsymbol{A}'$ の零空間における基底ベクトルの計算を例示する。行列 $\boldsymbol{A}'$ は

$$\boldsymbol{A}' = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

と表わされる。行列 $\boldsymbol{A}'$ の零空間を $\boldsymbol{x} = (x_1, x_2, \cdots, x_9)'$ とすると、$\boldsymbol{A}'\boldsymbol{x} = \mathbf{0}$ は

$$x_2 + x_3 + x_6 + x_8 + x_9 = 0$$

$$x_4 + x_6 + x_7 + x_8 + x_9 = 0$$

$$x_5 = 0$$

が成り立つ。したがって、

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} = \begin{bmatrix} x_1 \\ -x_3 - x_6 - x_8 - x_9 \\ x_3 \\ -x_6 - x_7 - x_8 - x_9 \\ 0 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_6 \begin{bmatrix} 0 \\ -1 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_7 \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_8 \begin{bmatrix} 0 \\ -1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + x_9 \begin{bmatrix} 0 \\ -1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

となり、行列 $\boldsymbol{A}'$ の零空間における $6\,(= 3^2 - 2^2 + 1)$ 個の基底ベクトルが求められた。

10.3 GIBBS SAMPLING FOR MODEL $\mathrm{M_{t,\alpha}}$

個体が潜在履歴 $\boldsymbol{\omega}_i$ をもつ確率：

$$\pi_i = \prod_{t=1}^{T} p_t^{\mathbf{I}(\omega_{i,t}>0)} (1-p_t)^{\mathbf{I}(\omega_{i,t}=0)} \alpha^{\mathbf{I}(\omega_{i,t}=1)} (1-\alpha)^{\mathbf{I}(\omega_{i,t}=2)}$$

個体数 N において、潜在履歴頻度 $\boldsymbol{X} = (x_1, x_2, \cdots, x_{3^T})'$ が起こる確率：

$$[\boldsymbol{X} \mid N, p_1, p_2, \cdots, p_T, \alpha] = \left\{ \frac{N!}{\prod x_i!} \prod \pi_i^{x_i} \right\} \mathbf{I}\left(\sum x_i = N \right)$$

潜在履歴頻度を考慮したときの、記録履歴頻度 $\boldsymbol{f}^+$ が起こる確率：

$$[\boldsymbol{f}^+ \mid \boldsymbol{X}, N, p_1, p_2, \cdots, p_T, \alpha] = \mathbf{I}(\boldsymbol{f}^+ = \mathbf{A}'\mathbf{X})$$

全条件付き分布は、事前分布を考慮すると

$$[\boldsymbol{f}^+ \mid \boldsymbol{X}, N, p_1, p_2, \cdots, p_T, \alpha]\,[\boldsymbol{X} \mid N, p_1, p_2, \cdots, p_T, \alpha]\,[N]\,[p_1]\,[p_2] \cdots [p_T]\,[\alpha]$$

に比例する分布である。

Prior and Full Conditionals for p_t *and* α

p_t と α の事前分布がベータ分布 $p_t \sim \mathrm{Be}({a_0}^t, {\mathrm{b}_0}^{\mathrm{t}})$ および $\alpha \sim \mathrm{Be}({a_0}^{\alpha}, {\mathrm{b}_0}^{\alpha})$ に従うとき、全条件付き分布もまたベータ分布 $p_t \sim \mathrm{Be}(a^t, \mathrm{b}^{\mathrm{t}})$ および $\alpha \sim \mathrm{Be}(a^{\alpha}, \mathrm{b}^{\alpha})$ に従う。

$$a^t = {a_0}^t + \sum_i x_i\, \mathbf{I}(\omega_{i,t} > 0)$$

$$b^t = {b_0}^t + \sum_i x_i\, \mathbf{I}(\omega_{i,t} = 0)$$

$$a^{\alpha} = {a_0}^{\alpha} + \sum_i \sum_t x_i\, \mathbf{I}(\omega_{i,t} > 0)$$

$$b^{\alpha} = {b_0}^{\alpha} + \sum_i \sum_t x_i\, \mathbf{I}(\omega_{i,t} = 0)$$

Prior and Full Conditionals for $\boldsymbol{X}$ *and* N

$\boldsymbol{X}$ を固定すると新しい N の値が標本抽出できないため、$(\boldsymbol{X}, N)$ の一組として Gibbs sampling を行う。この連結全条件付分布は、$\mathrm{g}(N)$ を N の事前分布とすると

$$[\boldsymbol{X}, N \mid \cdot] \propto \left\{ \frac{N!}{\prod x_i!} \prod \pi_i^{x_i} \right\} \mathbf{I}(\boldsymbol{f}^+ = \mathbf{A}'\mathbf{X})\, \mathbf{I}\left(\sum x_i = N \right)\, \mathrm{g}(N)$$

と表現される。 N に関する事前知識がない場合、適当な大きな値 M を与えて、事前分布に離散的な一様分布 $\{1,2,\cdots,M\}$ を用いるのが妥当かもしれない。有限な範囲をもたない非正則な一様分布 $[N] \propto 1/N$ も事後分布は正則となるので、事前分布として用いることができる。そこで、複数の事前分布を使って解析結果の感度を評価することが勧められる。

Gibbs Sampling

Gibbs sampling は次のように行われる。

Step 1: 妥当な潜在履歴頻度の集合から初期値 $\boldsymbol{X}_{\text{curr}}$ を発生させる。

Step 2: $\boldsymbol{X}_{\text{curr}}$ から a^{α}, b^{α} と a^{t}, b^{t} (for $t = 1,2,\cdots,T$) を計算する。

Step 3: p_t と α を標本抽出する。

Step 4: $k = 0$ とおく。

Step 5: $k = k + 1$ とする。0 を除いた離散的な一様分布 $\{-D_k,\cdots -1,1,\cdots,D_k\}$ から c_k を標本抽出し、潜在履歴頻度の候補 $\boldsymbol{X}_{\text{cand}} = \boldsymbol{X}_{\text{curr}} + c_k\,\boldsymbol{b}_k$ を発生させる。ここで $\boldsymbol{b}_k$ は $\text{null}(\boldsymbol{A})$ の基底ベクトル、D_k は調節パラメータである。

Step 6: $r = \min\left\{\dfrac{[\boldsymbol{X}_{\text{cand}}|\cdot]}{[\boldsymbol{X}_{\text{curr}}|\cdot]}, 1\right\}$ を計算し、r の確率で候補値を採択する（i.e. $\boldsymbol{X}_{\text{curr}} = \boldsymbol{X}_{\text{cand}}$）。

Step 7: $3^T - 2^T + 1$ 個の基底ベクトルすべてで、Step 5 ~ Step 6 を繰り返す。

Step 8: Step 2 ~ Step 7 を十分に多く繰り返す。

10.4 AN IMPLEMENTATION OF MODEL $\mathrm{M}_{t,\alpha}$

調査回数 $T = 5$、個体数 $N = 400$、捕獲確率 $p_1 = 0.3, p_2 = 0.4, p_3 = 0.5, p_4 = 0.6, p_5 = 0.7$、正識別率 $\alpha = 0.9$ から記録履歴頻度を発生させ、シミュレーションによる Model $\mathrm{M}_{t,\alpha}$ の個体数推定を行った。

まず、Model M_t で平坦な事前分布を用いて推定を行ったところ、個体数 N は 25% ほどの過大評価がされた（事後分布の中央値 = 511 , 95% HPDI = [495, 525] ）。Model $\mathrm{M}_{t,\alpha}$ による推定は 110000 回の標本抽出と最初の 10000 回を焼き捨てにより行われた。最初の 5000 回は D_k の調節期間としても用いた。$\delta_k = 1$ から始め、k^{th} の基底ベクトルの新しい候補が採択された場合には δ_k を 0.95 倍し、採択されなかった場合には 1/0.95 倍する†。Step 5 の D_k には、δ_k より大きい整数の中で最も近いものを当てはめた。

Model $\mathrm{M}_{t,\alpha}$ の結果は満足のいくものであった（Table 10.3）。

Markov 連鎖はかなり長い自己相関を示しており（Figure 10.1）、これは Metropolis-Hastings の標本抽出における低い移動率（標本抽出された値の変化が小さいということ？）に原因があるようだ。調節された 212 個の $\mathrm{D_k}$ の多くが 3 以下の値を示し、1 個（頻度 0 の潜在履歴を増やす基底ベクトル）の $\mathrm{D_k}$ だけが 12 となった。さらに、この低い移動率は潜在履歴の多くが低い頻度をもつことによるようだ。243 種類の潜在履歴のうち、145 種類が頻度 0 となり、215 種類が頻度 3 以下となった。残り 28 種類の潜在履歴に注目すると実際のものとよく一致していた（Figure 10.2）。

---

† つまり、基底ベクトルの新しい候補値が採択されるときは十分大きな事前分布から候補地が選ばれているとして δ_k の値を小さく、逆に基底ベクトルの新しい候補値が採択されないときは事前分布の大きさが十分でないとして δ_k の値を大きくしている。

10.5 EXTENSIONS

標識再捕獲データの多くはベルヌーイ試行に従う事象の履歴として記述される。事象の履歴が完全に観察される場合、モデルのパラメータ推定は容易だ。しかし、実際のデータの多くは、事象の履歴というよりは、それらが複合した捕獲履歴であり、その要素となる事象の多くが直接観察されない。こうした複数の事象からなる履歴は、容易な推定を許さない尤度関数を導く。本章のモデル手法は、そのような複合的な履歴が観察される多くの場合に応用できる。多項分布に従う潜在的な確率変数 $\boldsymbol{X}$ を観察できない事象の履歴を記述し、変換により頻度ベクトル $\boldsymbol{f}$ に集約されたものとして考えることができる。また、本章で発達させた Gibbs sampling の方法も他の標識再捕獲モデルの多くに応用できるだろう。

Model $M_{t,\alpha}$ とその他のモデルの違いは、行列 $\boldsymbol{A}$ の行に含まれる 1 の数が複数になることだ。ひとつの潜在履歴頻度が複数の記録履歴頻度を生じさせる。それゆえ Model $M_{t,\alpha}$ を解析するうえで、潜在的な事象を扱う多項分布の構造に対しての知識が必要となる。こうした理解は、必ずしも他のモデルでは必要にならないが、解析に対して明瞭な概念や多目的な枠組みを与えてくれるだろう。

PANEL 10.1

```
model {
    for ( t in 1:T ) { p[ t ] ~ dunif ( 0,1 ) }              # 捕獲確率 P_t の事前分布
    for ( j in 1:Cells ) {
        for ( t in 1:T ) {                                   # セル確率の計算の一部 p_t^ω_j,t (1 − p_t)^1−ω_j,t
            c[ j,t ] <- pow( p[ t ] , omegas[ j,t ] ) *
                    pow( 1 - p[ t ] , 1 - omegas[ j,t ] )
        }
        pi[ j ] <- prod( c[ j,1:T ] )                        # セル確率 π_j : p_t^ω_j,t (1 − p_t)^1−ω_j,t の積の計算
    }
    for ( j in 2:Cells ) {
        pi.obs[ j ] <- pi[ j ] / ( 1 – pi[ 1 ] )             # 観察履歴のセル確率 π_j^+
    }
    n <- sum( f [ 2: Cells] )                                # 観察履歴から捕獲された個体数 n を計算
    f [ 2: Cells] ~ dmulti ( pi.obs[ 2:Cells ] , n )         # 観察履歴 f^+ : ここで尤度を計算
    pn <- 1 – pi [ 1 ]
    n ~ dbin ( pn , N )                                      # 捕獲された個体数と捕獲確率から個体数 N を推定
    cN ~ dunif (n , 10000 )                                  # 個体群に含まれる個体数 N の事前分布
    N <- round (cN)                                          # 個体数 N を自然数に
}
```