Bayesian Inference: With ecological applications

(W.A. Link and R.J. Barker. 2010. Academic Press)

第 7章 Multimodel Inference

この章前半 (pp.127–139) の担当: 久保拓弥　kubo@ees.hokudai.ac.jp

paragraph #1　生態学的なプロセスの推定はほとんど必然的に model based である，つまりモデルだのみ

……データ集めの段階で，どんなにがんばってもね．そのモデルには，研究対象である生態学的なプロセスをあらわすコンポーネントがあり，そしてコンポーネントはデータと対応してないといけない．コンポーネントにはいろいろな種類のものが含まれるだろうけど，データにもとづいて推定される未知パラメーターをのぞいて，完全に指定されてなければならない．

paragraph #2　未知パラメーターの推定できました，でオワってしまうことが多い

しかしながら，その結果はあるモデルのもとでという条件つきのものであり，モデルを変えたらまたハナシが変わるでしょう．

paragraph #3　実際のところ，ひとつの現象を説明できそうなモデルはいろいろある

モデルにどういうコンポーネント /* (久保) たとえば説明変数とか */ をいれたらいいのか不確定であり，ある特定のモデルだけによる推定というのは，このあたりの不確定性を隠匿している．

paragraph #4　モデルが複雑なときにはモデル選びは実用的に重要だ

「興味あるパラメーター」(parameters of interest) の推定を正確にするために，モデル中に「あまり興味ないパラメーター」(nuisance parameters) /* (久保) たとえば，個体差とか random effects 的なやつ……いわゆる fixed effects 的なものも nuisance になりうるけれど */ をいれるんだけど，よけーな nuisance parameters はもちろん，nuisance じゃないパラメーター数も増やしたくない．

paragraph #5　ある一個のモデルを選んで推定するのではなく，推定過程の一部にモデル不確定性を含めることができればよいのだが……

そして，モデル内・モデル間の不確定性を反映した推定結果を得たい．たとえば K 個のモデル候補があったとして，生残率 $\hat{\phi}_k$ とその標準誤差 $s(\hat{\phi}_k)$ を推定したい /* (久保) この $\hat{\phi}_k$ は最尤推定値とかでしょう */ ．Buckland *et al.*(1997) は $\sum_{k=1}^{K} w_k = 1$ と規格化された「重み」を使って，こんなふうに combining model specific estimate

$$\tilde{\phi} = \sum_{k=1}^{K} w_k \hat{\phi}_k \tag{7.1}$$

あるいは composite measure of uncertainty

$$s(\tilde{\phi}) = \sum_{k=1}^{K} w_k \sqrt{s(\hat{\phi}_k)^2 + (\hat{\phi}_k - \tilde{\phi})^2}$$

を提案してみた．(てきとうな数値計算例) ……これにはモデル内・モデル間のばらつきが含まれている．

paragraph #6 複数モデル推定の問題は「モデル選択」と「モデル重みづけ」

モデル選択はいくつかの候補の中から一番よいものを選ぶ．選択のためには推定が必要．モデル重みづけは，モデルごとの推定結果に重みづけをして混ぜるんだけど，そのモデルのたしからしさとか不確定性がそこに加味される．

paragraph #7 しかし，現在のところ（そしてたぶん将来も）誰もが納得してくれる方法はない……

ベイズでどうこうしよう，というアプローチも同様.

paragraph #8 だけど Bayesian Multimodel Inference (BMI) って良いんじゃないかな?

というのも，第一章で述べたように，ベイズ的な方法にはいろいろ良いところがあって，BMI はその 「自然な」拡張になっている．あるモデルの事前確率を設定し，推定された事後モデル確率でモデルを組みあわせられる．複数のモデルの中から一個モデルを選びたい場合でも，その選択の規準を計算できる．

paragraph #9 この章であつかうことは……

ベイズ因子 (Bayes factor)，ベイズ情報量規準 (Bayesian information criterion; BIC)，Deviance ifnromation criterion (DIC)．可逆ジャンプマルコフ連鎖モンテカルロ法 (Rerversible jump Marcov chain Monte Carlo) `/*` (**久保**) **こんな訳語でいいのか**? `*/` ，そして BUGS による BMI の簡単な実装など．

paragraph #10 BMI で難しいことのひとつは……

あいまいな事前分布 (vague prior) `/*` (**久保**) **いわゆる無情報事前分布のたぐい** …… 7.2.5 **項に登場** `*/` の選択が難問である．モデルが一個のときは，どういうあいまい事前分布でもよいのだけど，複数モデルの場合はそう簡単ではない． `/*` (**久保**) **あとからまた議論** `*/`

paragraph #11 ちかごろ生態学では AIC ってのがハヤリだけど……

この章では，このあたりもベイズと関連づけて議論しよう.

7.1 The BMI Model

paragraph #1 BMI ってのはホントにベイズ統計モデルそのまま!

変数はすべて確率変数，事後分布を推定.

paragraph #2 BMI では，まず $Model$ なる確率変数を考えて……

これは「自然」(Nature) が **Fig. 7.1** みたいなモデルがたくさん (K 個) 入ったバケツからある

モデルを選ぶ確率．さらに「自然」はパラメーターを事前分布からパラメーターを選び，選んだモデルで $Data$ を生成している．

paragraph #3　統計学的な推定のほとんどでは……

「自然」がどのモデルを選んだかを知ってる，フリをしている．それに対して，BMI はモデルの不確定性を認知している．

paragraph #4　$Model$ は多項分布の確率変数だと思って……

それぞれのモデルが選ばれる確率を $\pi_1, \pi_2, \cdots, \pi_K$ としてみる．モデルをひいきしないなら，$\pi_k = 1/K$，等確率．あるいは事前の信念とやらを反映させてもよい．最節約的なモデルが良いとかね．

Objections! (異議あり!)

paragraph #5　異議 1: 自然はバケツからモデルを選ばない!

まあ，便利だからいいじゃない．

paragraph #6　異議 2: そのバケツの中に真のモデルがなかったら?

それは考えてもしょうがないので，$Model\ k \in \{1, 2, \cdots, K\}$ の中に真なモデルがあると思って計算することにしよう．

paragraph #7　バケツモデルそれ自体ひとつのモデルだし

これがおとしどころ，ということで．

7.1.1 Example: BMI for Two Fully Specified Models

paragraph #1　幾何分布とポアソン分布

/* (久保) いきなりふたつの統計モデリングのハナシです */ 幾何分布 (geometric distribution) の確率密度関数 (PDF) は $p(1-p)^y$ で，ポアソン分布 (Poisson distribution) は $\exp(-\lambda)\lambda^y/y!$．どちらも $y = 0, 1, 2, \cdots$ な値をとる．

paragraph #2　幾何分布モデル vs ポアソン分布モデル

てもとに $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_5\}$ というデータがあり，幾何分布モデルもしくはポアソン分布モデルが生成した．M は「自然」が選ぶモデルをあらわす名義変数な確率変数．幾何分布モデルの尤度は，

$$\Pr(\mathbf{Y}|M_1, p) = \prod_{k=1}^{5} p(1-p)^{Y_i} = p^5(1-p)^{5\bar{Y}} \tag{7.2}$$

となり，この式の中の $\bar{Y} = \frac{1}{5}\sum_{i=1}^{5} Y_i$．ポアソン分布モデルは

$$\Pr(\mathbf{Y}|M_2, \lambda) = \prod_{k=1}^{5} \frac{\exp(-\lambda)\lambda^{Y_i}}{Y_i!} = \frac{\exp(-5\lambda)\lambda^{5\bar{Y}}}{\prod_{i=1}^{5} Y_i!} \tag{7.3}$$

となる．M_k と書いてるのはホントは $M = M_k$ だけど，上のように書くのがラクなのでそうする．

paragraph #3 簡単のため真の平均 (pupulation mean) は 3，だと知ってる

つまり $p = 1/4$ /* (久保) 幾何分布の平均は $p/(1-p)$ なので */ または，$\lambda = 3$．**Fig. 7.2** の確率密度関数を参照．以下では $\Pr(\mathbf{Y}|M_1, p)$ を $\Pr(\mathbf{Y}|M_1)$ と書く．

paragraph #4 M_1 である事後確率は

π が「M_1 が選ばれる事前確率」だとすると

$$\Pr(M_1|\mathbf{Y}) = \frac{\pi \Pr(\mathbf{Y}|M_1)}{\pi \Pr(\mathbf{Y}|M_1) + (1-\pi)\Pr(\mathbf{Y}|M_2)} \tag{7.4}$$

となる．

paragraph #5 さて，たとえばデータが $\mathbf{Y} = \{0, 1, 2, 3, 8\}$ だとすると

平均は 2.8 だけど分散は 9.7．ポアソン分布は平均と分散が等しいので，このデータへのあてはまりはよくなさそう．幾何分布だと良さそう /* (久保) 幾何分布は 分散 = 平均 ^{2} なので */ ．幾何分布が選ばれる事前確率を $\pi = 0.5$ とすると事後確率 $\Pr(M_1|\mathbf{Y})$ は 0.852 になる．つまりオッズ (odds) が (0.5 : 0.5) つまり (1 : 1) から (0.852 : 0.148) つまり (5.75 : 1) に変わった．

paragraph #6 こういうオッズの変化は BMI のまとめとして有用

式 (7.4) と同じく，M_2 に関してはこうなる．

$$\Pr(M_2|\mathbf{Y}) = \frac{(1-\pi)\Pr(\mathbf{Y}|M_2)}{\pi \Pr(\mathbf{Y}|M_1) + (1-\pi)\Pr(\mathbf{Y}|M_2)} \tag{7.5}$$

式 (7.4) と (7.5) の両辺をわると

$$\frac{\Pr(M_1|\mathbf{Y})}{\Pr(M_2|\mathbf{Y})} = \frac{\pi}{1-\pi} \times \frac{\Pr(\mathbf{Y}|M_1)}{\Pr(\mathbf{Y}|M_2)} \tag{7.6}$$

事後モデルオッズ $\frac{\Pr(M_1|\mathbf{Y})}{\Pr(M_2|\mathbf{Y})}$ は事前モデルオッズ $\frac{\pi}{1-\pi}$ をデータの相対確率 $\frac{\Pr(\mathbf{Y}|M_1)}{\Pr(\mathbf{Y}|M_2)}$ /* (久保) ← これは尤度比 */ でスケイリングしたものになっている．

paragraph #7 式 (7.6) がベイズ因子 (Bayes factor) の定義

データ $\mathbf{Y}$ が定まったものであるとき，上の式をコトバでいいかえると，

$$\text{Posterior model odds} = \text{Prior model odds} \times \text{Bayes factor}$$

この例題での BF は 5.75.

7.1.2 Example: BMI with Unknown Parameters

paragraph #1 今度は p や λ がわからない場合

この場合，p や λ の事前分布が必要になる．

paragraph #2　平均確率を評価する

事前分布で重みづけをした平均値.

paragraph #3　事前分布 $g(p)$ と $h(\lambda)$ を導入

尤度はこのように定義される．

$$\Pr(\mathbf{Y}|M_1) = \int p^5(1-p)^{5\bar{Y}} g(p)dp \tag{7.7}$$

$$\Pr(\mathbf{Y}|M_2) = \int \frac{\exp(-5\lambda)\lambda^{5\bar{Y}}}{\prod_{i=1}^{5} Y_i!} h(\lambda)d\lambda \tag{7.8}$$

paragraph #4　一様分布な事前分布をつかう

事前分布について考えるのはめんどうなので，あとまわし．λ の事前分布は $h(\lambda) = U(0,T)$ /* (久保) 0 から T までの一様分布 */ とする．p.21 の変数変換 (change of variables) 定理から，p の事前分布は

$$g(p) = \frac{1}{Tp^2}$$

となる $(1/(T+1) < p < 1)$． /* (久保) 幾何分布の平均を m とする．$m = (1-p)/p$ なので $dm/dp = 1/p^2$．m の事前分布 $f(m) = 1/T$ (範囲 $0 < m < T$) とすると，この m を p に変数変換した確率密度関数は $g(p) = \frac{1}{T} \times \frac{dm}{dp} = \frac{1}{Tp^2}$ となる． */

paragraph #5　データが $\mathbf{Y} = \{0,1,2,3,8\}$ だとすると

幾何分布モデルが良い，Bayes factor (BF) が 13.84 になるんで /* (久保) これって数値積分したのか? */ ．なぜ平均が未知のほうがよいのだろうか? 平均値が既知の場合の BF は 5.75 だったのに.

paragraph #6　平均が標本平均に一致する分布で BF 最小

Fig. 7.3 でそのように示されている．平均 2.8 で BF は 5.61 /* (久保) パラメーターの事前確率が異なるので BF = 5.75 にならない */ ．ここでモデル間の格差は最小．ところが，ここからずれた平均のときほど，ポアソン分布からの逸脱は相対的に大きくなるので，幾何分布が有利になる．

7.2 Bayes Factors

paragraph #1　Bayes factor で比較できるモデルは……

前の節でみたように，ネストしてる /* (久保) つまり M_1 が M_2 を簡単化したものである */ 必要がない．この節は Bayes factor のいろいろな性質を見る．

7.2.1 Bayes Factors and Likelihood Ratio Statistics

paragraph #1 Bayes factor は尤度比の一種

いま K 個のモデルがあって，モデル M_k の未知パラメーターは θ_k とする．かぎかっこ表記 (bracket notation) で書くとデータ $\mathbf{Y}$ の確率分布は $[\mathbf{Y}|M_k, \theta_k]$ で，パラメーターの事前分布は $[\theta_k|M_k]$ となる．データ $\mathbf{Y}$ が定まっているとき，$[\mathbf{Y}|M, \theta]$ はモデルとパラメーターの同時尤度 (joint likelihood) であり，θ で積分するとモデル M の周辺尤度 (marginal likelihood) が得られる．

$$[\mathbf{Y}|M] = \int [\mathbf{Y}, \theta|M] d\theta = \int [\mathbf{Y}|\theta, M][\theta|M] d\theta \tag{7.9}$$

BF はこの周辺尤度の比．

$$\mathrm{BF}_{i,j} = \frac{[\mathbf{Y}|M_i]}{[\mathbf{Y}|M_j]}$$

頻度主義的な考えかた (frequentist) の尤度比検定統計量に相当する．ただし尤度比検定では θ の最尤推定値を使うが，

$$\mathrm{LR}_{i,j} = \frac{[\mathbf{Y}|\hat{\theta}_i, M_i]}{[\mathbf{Y}|\hat{\theta}_j, M_j]}$$

BF は θ に関する平均である．

7.2.2 Bayes Factors are Multipliers of Odds

/* (久保) この項ではひたすら BF の書きかえやってるだけ */

paragraph #1 BF は事前オッズの比例定数

K 個のモデルがあるとき

$$\Pr(M_i|\mathbf{Y}) = \frac{[\mathbf{Y}|M_i]\pi_i}{\sum_{k=1}^{K}[\mathbf{Y}|M_k]\pi_k} \tag{7.10}$$

となるので，ここから事後確率オッズは

$$\frac{\Pr(M_i|\mathbf{Y})}{\Pr(M_j|\mathbf{Y})} = \frac{[\mathbf{Y}|M_i]}{[\mathbf{Y}|M_j]} \times \frac{\pi_i}{\pi_j} = \mathrm{BF}_{i,j} \times \frac{\pi_i}{\pi_j}$$

となり式 (7.6) を一般化したものになる．$\frac{\pi_i}{\pi_j}$ が 1 なら事後確率オッズは BF と同じ．

paragraph #2 Bayes factor は事前モデル確率 π_i に依存しない

ということで，こういう関係も成立する．

$$\mathrm{BF}_{1,3} = \mathrm{BF}_{1,2}\mathrm{BF}_{2,3} \tag{7.11}$$

paragraph #3 式 (7.10) は BF でも書ける

$$\Pr(M_i|\mathbf{Y}) = \frac{\mathrm{BF}_{i,1}}{\sum_{k=1}^{K}\mathrm{BF}_{k,1}\pi_k} \tag{7.12}$$

7.2.3 Updating Bayes Factors

paragraph #1　追加データで Bayes factor のアップデイト

……が BF のウリである．最初のデータ $\mathbf{Y}_1$ をとったときの BF は

$$\mathrm{BF}_{i,j}(\mathbf{Y}_1) = \frac{[\mathbf{Y}_1|M_i]}{[\mathbf{Y}_1|M_j]} = \frac{\int[\mathbf{Y}_1|M_i,\theta][\theta|M_i]d\theta}{\int[\mathbf{Y}_1|M_j,\theta][\theta|M_j]d\theta}$$

となり，次のデータ $\mathbf{Y}_2$ をとったときの BF は

$$\mathrm{BF}_{i,j}(\mathbf{Y}_2|\mathbf{Y}_1) = \frac{[\mathbf{Y}_2|\mathbf{Y}_1,M_i]}{[\mathbf{Y}_2|\mathbf{Y}_1,M_j]} = \frac{\int[\mathbf{Y}_2|\mathbf{Y}_1,M_i,\theta][\theta|\mathbf{Y}_1,M_i]d\theta}{\int[\mathbf{Y}_2|\mathbf{Y}_1,M_j,\theta][\theta|\mathbf{Y}_1,M_j]d\theta}$$

となる．こちらではパラメーター θ の事前分布に $\mathbf{Y}_1$ が入っている /* (久保) このあたり，説明が明瞭でないような気がするんだけど，$[\theta|\mathbf{Y}_1,M_i]$ とは θ の事後分布のことだろう．似たような考え方は 7.2.5 節最後の posterior Bayesian factor の説明にも登場する．事後分布を使ったあとづけ計算は，たしかにあれこれと便利ではありますけどねえ…… */ ．このように定義すると，BF は

$$\begin{aligned}\mathrm{BF}_{i,j}(\mathbf{Y}_2,\mathbf{Y}_1) &\equiv \frac{[\mathbf{Y}_2,\mathbf{Y}_1|M_i]}{[\mathbf{Y}_2,\mathbf{Y}_1|M_j]} = \frac{[\mathbf{Y}_2|\mathbf{Y}_1,M_i]}{[\mathbf{Y}_2|\mathbf{Y}_1,M_j]} \times \frac{[\mathbf{Y}_1|M_i]}{[\mathbf{Y}_1|M_j]} \\ &= \mathrm{BF}_{i,j}(\mathbf{Y}_2|\mathbf{Y}_1) \times \mathrm{BF}_{i,j}(\mathbf{Y}_1)\end{aligned}$$

このように分割できる．頻度主義的な検定にはこれに対応するものはない． /* (久保) ようするに，もともとの BF である $\mathrm{BF}_{i,j}(\mathbf{Y}_1)$ が新データ $\mathbf{Y}_2$ によって，$\mathrm{BF}_{i,j}(\mathbf{Y}_2|\mathbf{Y}_1)$ 倍の改善 (改悪) されました……というだけのことで，何かの役にたつのかな? ……MCMC 計算時間の短縮とか? */

7.2.4 Bayes Factors as Measures of Relative Support

paragraph #1　$\mathrm{BF}_{i,j}$ ってどれぐらい大きければいいの?

$\mathrm{BF}_{i,j} = 5$ ってどういう意味?

paragraph #2　これは事前モデル確率 π に対する事後モデル確率の依存性が，BF によってどう変わるかを調べればよさそう

式 (7.4) は以下のように書きかえられるので,

$$\Pr(M_1|\mathbf{Y}) = \frac{\mathrm{BF}_{1,2}\pi}{\mathrm{BF}_{1,2}\pi + (1-\pi)} \tag{7.13}$$

$$\Pr(M_1|\mathbf{Y}) \geq p_0 \quad \text{if and only if} \quad \pi \geq \frac{p_0}{p_0 + \mathrm{BF}_{1,2}(1-p_0)}$$

たとえば，$\mathrm{BF} = 50$ のときには，$\Pr(M_1|\mathbf{Y}) \geq 0.9$ となるためには $\pi \geq 0.16$ でなければならない.

paragraph #3 あるいは Fig. 7.4 を見よ

paragraph #4 Harold Jeffreys の BF 分類: Table 7.1

似たような分類はほかにもいろいろある /* (久保) うーむ，どうやって決めてるんだろう? */ . 式 (7.13) に対して $\pi = 0.5$ とすると，モデル M_1 の事後確率は,

$$\Pr(M_1|M_1 \text{ or } M_2, \mathbf{Y}) = \frac{\mathrm{BF}_{1,2}}{1 + \mathrm{BF}_{1,2}}$$

となり，これは頻度主義な検定のてきとーなる危険率 α 設定より良い.

7.2.5 Problems with Vague Priors on Parameters

paragraph #1 事前分布が vague だと，事後分布はデータで決まる

/* (久保) 前にも書いたけど，vague な事前分布とは無情報事前分布 */ たとえば，$[X|\mu] = N(\mu, 1)$ で $[\mu] = N(\mu_0, \sigma^2)$ としよう．パラメーター μ の事後分布は

$$[\mu|X] = N\left(\frac{1}{1+\sigma^2}\mu_0 + \frac{\sigma^2}{1+\sigma^2}X, \frac{\sigma^2}{1+\sigma^2}\right)$$

となり，$\sigma \to \infty$ とすると，これは μ の尤度関数に近づく．ということで，σ の値を十分に大きくしてやると— つまり μ について知識がないといったことなのだが— 事後分布は尤度で決まるようになり，事前分布の影響がなくなる.

paragraph #2 $\sigma = \infty$ な事前分布は improper だよね

つまり確率分布になってない，ということ．こういう場合は σ を十分に大きくすればよくて，$\sigma = 100$ だろうが $\sigma = 10^6$ だろうがほとんど同じ事後分布になる.

paragraph #3 こういう事前分布は複数モデルあつかうときにめんどう

モデル一個の推定なら上のような事前分布でよいのだけど．とくに，モデルごとにパラメーター数が異なる場合はたいへん．ここまでの正規分布モデルの例でいうと,

- Model1: $[X|M = 1] = N(0, 1)$ /* (久保) なぜか $M = M_1$ 記法をやめてしまったようだ */
- Model2: 事前分布を $[\mu] = N(\mu_0, \sigma)$ としたので $[X|M = 2] = N(\mu_0, 1)$

となっているときに，BF は

$$\mathrm{BF}_{1,2} = \cdots(\text{写経に疲れました})\cdots = \sqrt{1+\sigma^2}\exp\left(-\frac{1}{2}\left\{\frac{X^2\sigma^2 + 2X\mu_0 - \mu_0^2}{1+\sigma^2}\right\}\right)$$

これは $\sigma \to \infty$ とすると，$\mathrm{BF}_{1,2} \to \infty$ となってしまう．データ X に関係なくこうなってしまうので，常に Model 1 が良いということになる.

paragraph #4　Posterior Bayes factor (PBF)

これは事後平均尤度なんだけど，それを比較すれば良いのではと Aitkin (1991) が提案した．まずは，BF の評価に必要な，周辺分布の計算に立ちかえって，その定義を示すと /* (久保) 下の式は，単に θ を θ_M にしただけのもの */

$$[\mathbf{Y}|M] = \mathrm{E}_{[\theta_M|M]}([\mathbf{Y}|M,\theta_M]) = \int [\mathbf{Y}|M,\theta_M][\theta_M|M]d\theta_M$$

そして，Aitkin の提案は，上の式の $[\theta_M|M]$ を $[\theta_M|M,\mathbf{Y}]$ に置きかえてしまって……

$$\mathrm{E}_{[\theta_M|M,\mathbf{Y}]}([\mathbf{Y}|M,\theta_M]) = \int [\mathbf{Y}|M,\theta_M][\theta_M|M,\mathbf{Y}]d\theta_M$$

これを使って BF を評価せよ，というもの． /* (久保) このあたり，説明が明瞭では内容な気がするのだが……つまり，$[\theta_M|M,\mathbf{Y}]$ とは θ_M の事後分布のことで，「事後分布で尤度を重みづけした事後確率」が PBF ……これってデータの「二度づけ」だろ，と批判されてる． */

paragraph #5　PBF は同じデータを二回つかってるからダメ，と叩かれている

データにあうように選んだパラメーターで，重みを評価するとそれは過大評価になる．そういう欠点はあるけれど PBF は良さそう /* (久保) と著者はこだわっています */ ．

paragraph #6　Aitkin の別案: データわけろ

training 用データ /* (久保) つまりパラメーター推定用データ */ とモデルの重みづけ計算用データをわければよい，と提案． /* (久保) cross validation 的? */

paragraph #7　結論

マルチモデルな推定において，パラメーターの事前分布の選びかたが問題になる．このあたり，簡単ではない．先験的な好みを反映させぬよう，パラメーターの事前分布は選びたい．節約的なモデルが良いといったことは，モデルの事前確率に反映させるべきだ．このあたり 7.4.3 節の例題でまた議論する．