2003 年 1 月

最尤推定法メモ

苫小牧シウリザクラの父性判定結果を利用した統計モデリング

久保拓弥 kubo@ees.hokudai.ac.jp

http://hosho.ees.hokudai.ac.jp/~kubo/ml/

1　計算の目的と推定に用いるデータ

苫小牧研究林クレーンサイトと国有林緑のトンネルのシウリザクラの種子のマイクロサテライト DNA 解析から，花粉親の推定がなされた．そこで，この結果と同調査地で得られた他のデータと組み合わせて,「花粉親として成功する要因」の推定を試みた．さらに同じデータセットで，ある花序の結実が周辺に存在する花粉親の質・量にどう依存しているか，の統計モデルも構築してみた．

父親としての成功する要因の推定において，統計モデルで説明しようとする現象は「ある花序の父親が属するジェネットは何か」であり，これはマイクロサテライト DNA の解析結果である．この結果は決定論的でなく統計的な推測を含んでいる．しかしながら，今回の統計モデリングではこの推定結果が完全に正しい，と仮定している．

なお，ここではクレーンサイトと緑のトンネルの全個体が共通のパラメーターを持つものとして，推定計算を行った (なお，これ以降で「個体」と表記した場合にはシウリザクラの 1 ラメットをさす)．これはそれぞれの調査地 (とくにクレーンサイト) で母花序の空間分布の「偏り」が推定結果に影響を与えそうなためである．

第二の問題である「花序の結実を決める要因」の推定はある花序内での結実数を説明するモデリングを行った．これはクレーンサイトでのみ得られた観測データである．ただしこのデータセットでは結実数は得られているけれど，花序内の花数がわかっていないので，結実数そのものを被説明変数とした (いわゆる結実率がわからない)．推定においては，花序のサイズも説明変数として組み込んでいるので，花序サイズの違いを考慮した上で他の要因 (周囲に存在する花粉親の量) の効果を推定している．

このメモで使われる記号とシウリザクラ個体ごとに定まる説明変数は以下の通りである:

- i: 母花序のインデックス (母花序のジェネットも i であらわす)
- j, j': 花粉親 (父親) のジェネットのインデックス
- k, k': 花粉親 (父親) の個体のインデックス
- d_k: 花粉親個体 k の直径 (mm)
- r_{ik}: i (をもつ個体) と花粉親個体 k の間のユークリッド距離 (m)
- I_{ij}: i と j が同一ジェネットかどうかをあらわす変数 (同一なら 1 そうでなければ 0)
- s_k: k が単木性をあらわす変数 (単木なら 1 そうでなければ 0)
- l_i: 花序 i の長さ (mm)
- u_i: 花序 i が樹冠上部にあれば 1，そうでなければ 0 (上下不明な場合も 0)

今回の統計モデリングで花粉親候補とした個体は，上記のデータが得られる個体，すなわちクレーンサイト・緑のトンネルそれぞれの調査区画内に存在する個体のみである．調査区の外にいる個体については考慮していない．なお，シウリザクラの開花状況に関しては，父親推定と対応のつく観測データがないので，花粉親候補の全個体が開花している，と仮定した．

次節以降で推定されるパラメーターもここに列挙しておく (父親成功要因問題と花序内結実数問題の両方):

- β_0: 花粉親サイズ依存性
- β_1: 父母間距離依存性
- β_2: 自家 (不) 和合依存性
- β_3: 単木性依存性

- γ_0: (結実量モデルの定数項)
- γ_1: 花序サイズ依存性
- γ_2: 樹冠内の上下位置依存性
- γ_3: 周囲に存在する同ジェネット花粉親量依存性
- γ_4: 周囲に存在する他ジェネット花粉親量依存性

2　花粉親として成功する要因の推定と選択

まずある個体が花粉親として成功する確率を高める要因の推定と選択を行う．図 1 のように個体とジェネットをあらわす記号を導入する．ここでは結実してる花序

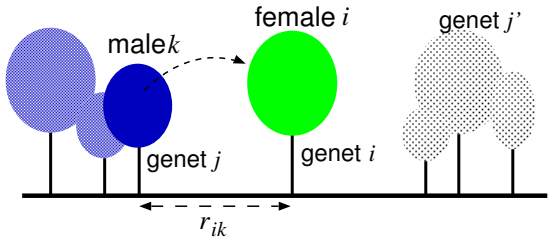

図 1: ジェネットと個体の関係

を i とあらわす．面倒なのでその花序が所属しているジェネットも i とあらわす．この花序 i の父親として特定されたジェネットを j とする．ジェネット j に含まれる個体を k であらわす．

各個体の「花粉親としての競争力」が個体のサイズや i-k 間の距離で決まると仮定する (花序 i の位置はその花序がついてる樹木個体の位置とした)．ここでこの競争力に比例して父親としてランダムに選ばれる “lottery model” を仮定すると，ある母親花序 i の父親が属するジェネットが j である確率，すなわち尤度 (likelihood) は

$$\begin{aligned} L_i &= \frac{\sum \text{ジェネット } j \text{ に属する個体 } k \text{ の花粉競争力}}{\sum \text{全ジェネットの個体 } k' \text{ の花粉競争力}} \\ &= \sum_{k \in \mathbf{G}_j} f_{ijk} \Big/ \sum_{\mathbf{G}_{j'} \in \mathbf{G}} \left(\sum_{k' \in \mathbf{G}_{j'}} f_{ij'k'} \right), \end{aligned}$$

として書ける．ここで $\mathbf{G}_j$ はジェネット j に属する全樹木個体の集合，$\mathbf{G}$ は父親となりうる全ジェネットの集合，f_{ijk} は上述の「花粉親としての競争力」である．競争力 f_{ijk} に関して以下のような仮定を導入した:

$$\begin{aligned} f_{ijk} &= (\text{個体 } k \text{ のサイズ依存項}) \times (i \text{ と } k \text{ 間の距離依存項}) \\ &\quad \times (\text{ジェネット依存項}). \end{aligned}$$

すなわち，f_{ijk} は父親候補の個体 k のサイズに依存する項，母親個体とのユークリッド距離に依存する項，母親と父親が同一ジェネットであるかどうか (部分的自家不和合性など) に依存する項それぞれに比例する，と仮定している．この具体的な関数型としては

$$f_{ijk} = d_k^{\beta_0} \exp(-\beta_1 r_{ik}) \exp(\beta_2 I_{ij}) \exp(\beta_3 s_k) \qquad (1)$$

と仮定した．個体 k ごとに異なる変数は以下のとおりである: d_k が個体 k の胸高直径 (DBH)，r_{ik} が母花序 i と父個体 k の間のユークリッド距離，I_{ij} が父個体 k が属してる j と母花序 i の遺伝的同等性をあらわし，

$$I_{ij} = \begin{cases} 0 & \text{if } i \text{ と } j \text{ が異なるジェネット} \\ 1 & \text{if } i \text{ と } j \text{ が同一ジェネット} \end{cases} \tag{2}$$

s_k は k が所属しているジェネットサイズが 1 本なら 1 (2 本より大きいなら 0)，と定義する．$\{\beta_0, \cdots, \beta_3\}$ は推定すべきパラメーターであり，それぞれサイズ依存性・父母距離依存性・自家和合性・ジェネットサイズ依存性をあらわす．もし β_2 が正の値であれば自家和合性かつ同一ジェネットの花粉親を積極的に選んでいることになり，逆に負の値であれば部分的自家不和合となる ($\beta_2 = -\infty$ で完全自家不和合)．β_3 は構成個体が 1 本であるようなジェネットの有利さ・不利さを示すことになる．

観測データ全体の尤度 L は各母親花序の尤度 L_i の積として表現される．これがこの推定問題の尤度方程式であり，

$$L(\beta_0, \cdots, \beta_3) = \prod_{i \in \mathbf{M}} L_i(\beta_0, \cdots, \beta_3),$$

ここで $\mathbf{M}$ は全母親花序の集合である．この L を最大化するようなパラメーター $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$ を Nelder-Mead 法で計算した (Perl スクリプトによる数値的最尤推定法)．両辺の対数をとると以下のような対数尤度方程式が得られる，

$$\log L(\beta_0, \cdots, \beta_3) = \sum_{i \in \mathbf{M}} \log \left(\sum_{k \in \mathbf{G}_j} f_{ijk} \right) - \sum_{i \in \mathbf{M}} \log \left(\sum_{\mathbf{G}_{j'} \in \mathbf{G}} \sum_{k' \in \mathbf{G}_{j'}} f_{ij'k'} \right).$$

さて，ここでパラメーター β_* の中から現象を説明するのに最小限必要なものを選択する操作 (モデル選択) が必要になる．パラメーターの全セット $\{\beta_0, \cdots, \beta_3\}$ から少なくとも 1 個以上を取りだす組み合わせ (選ばれたパラメーターがゼロでない値を取り，選ばれなかったものはゼロとおく) は $2^4 - 1 = 15$ とおりある．すべての β_* がゼロである場合は考えない．この 7 とおりの組み合わせすべてに関して最大化対数尤度を計算した．この最大化対数尤度と使用パラメーター数を同時に考慮したモデル選択基準 Akaike Information Criteria,

$$\text{AIC} = -2 \text{ (最大化対数尤度)} + 2 \text{ (パラメーター数)},$$

が最小となるパラメーターの組み合わせを選び出した．

3　母花序の結実量を決める要因の推定と選択

つぎに花序ごとの結実量を説明する要因を推定し選択する問題を考える．母花序 i の結実量 n_i (個) がポアソン分布に従う，と仮定する:

$$L_i = \frac{\lambda_i^{n_i} \exp(-\lambda_i)}{n_i!}, \tag{3}$$

ここで λ_i は花序 i の結実量の平均であり，その関数型は

$$\lambda_i = \exp(\gamma_0 + \gamma_1 l_i + \gamma_2 u_i + \gamma_3 P_i + \gamma_4 Q_i), \tag{4}$$

とする．花序 i ごとに特定される変数は，花序の長さ l_i，花序 i が樹冠上部にあれば u_i は 1 (下部もしくは上下不明な場合は 0)，花序 i の「周囲に存在する花粉親候補の量」P_i (i と同一ジェネット) と Q_i (i とは異なるジェネット) である (図 2)．つ

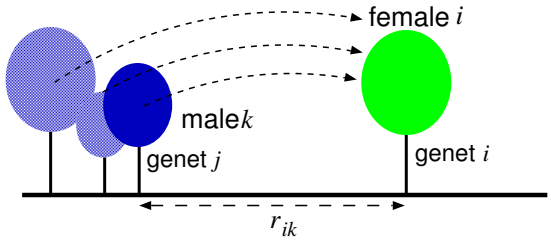

図 2: 「周囲に存在する花粉親候補の量」の概念

まり，この P_i と Q_i はは父親になりうるシウリザクラの個体数の重みつき平均であり，それぞれ式 (2) の I_{ij} を用いて，

$$\begin{aligned} P_i &= \sum_{j \in \mathbf{G}} I_{ij} g_{ij}, \\ Q_i &= \sum_{j \in \mathbf{G}} (1 - I_{ij}) g_{ij}, \end{aligned}$$

と定義される．この式からもわかるように，P_i には花序 i が存在する樹木個体自身も**含まれている**．ここで g_{ij} は母花序 i から見たジェネット j の「重み」を定義する関数である．これは j に含まれる父親候補個体のサイズと i から距離に依存する，と仮定した．その関数型は前節の式 (1) で定義された f_{ijk} から同一ジェネットの効果 (自家和合・不和合性) の項を除いたうえでジェネット j 全体で和をとり

$$g_{ij} = \sum_{k \in \mathbf{G}_j} d_k^{\beta_0} \exp(-\beta_1 r_{ik}),$$

とした．この関数にはサイズ依存性ならびに父母距離依存性の $\{\beta_0, \beta_1\}$ が含まれている．今回の推定計算では，前節の方法で得られた推定値 $\{\hat{\beta}_0, \hat{\beta}_1\}$ を代入して g_{ij} を計算した．

未知のパラメーター $\{\gamma_0, \cdots, \gamma_4\}$ の推定値を計算する尤度方程式は，L_i を全母花序に関して積をとったものとなる

$$L(\gamma_0, \cdots, \gamma_4) = \prod_{i \in \mathbf{M}} L_i(\gamma_0, \cdots, \gamma_4)$$

ところで式 (3) と式 (4) から，この推定は一般化線形モデル (ポアソン分布，link 関数は log) であることがわかる．そこで推定計算には R の `glm()` 関数を使い，モデル選択には `stepAIC()` 関数を用いた．また式 (4) の定数項 (γ_0) をランダム変量とみる一般化線形混合モデルによる推定も行った．これは花序ごとに γ_3 が変動する (正規乱数になっている) と仮定したモデルであり，R の `glmm()` 関数で推定値を評価した．

(とりあえずここまで)