確率分割の標本と予測量: 生態学への応用

2013-11-27/29 金沢大学

慶應義塾大学　渋谷　政昭

要約　　ピットマン確率分割の条件付き確率分割 $S \in \mathcal{P}_{n,k}|S \in \mathcal{P}_{\nu,\kappa}, n < \nu$ からランダムにサブサンプルを取り出す逆過程を調べ, $S \in \mathcal{P}_{n,k}$ による (ν, κ) の予測を行う. 関連する確率標本, 予測量にも触れる.

1　まえがき

生態学調査データと数の確率分割　　生態学調査の基本データは観測・捕獲した種 C_i の個体数 c_i, $1 \le i \le k$, である. $n := \sum_{i=1}^{k} c_i$, k, はそれぞれ, 観測・捕獲した個体総数, 種の数である. 伝統的には種 C_i の生存確率 を推定するが, 観測・捕獲されなかった種の確率が問題となる. 具体的には, 観測・捕獲を継続して n を増加したときに k がどう変わるか, n の異なるデータをどのように比較するか, などが課題である. ここでは $\{c_1, \ldots, c_k\}$ を n の確率分割とみなしてこの課題に挑戦する.

$\mathbb{N}$ を自然数の全体, 可能な分割の集合を

$$\mathcal{P}_{n,k} := \{\{c_1, \ldots, c_k\};\ c_i \in \mathbb{N}, 1 \le i \le k, \sum_{i=1}^{k} c_i = n\}, \quad \mathcal{P}_n := \bigcup_{k=1}^{n} \mathcal{P}_{n,k},$$

で表す. $\mathcal{P}_n$ の上の確率測度の系列 $\Pi_n, n = 1, 2, \ldots$ が**確率分割** (random partition) である.

確率分割では分類 C_i の順序を無視して, 下降順序統計量 $(c_1, \ldots, c_k)\downarrow$, あるいは同じ大きさの c_i の数

$$s_j := \sum_{i=1}^{n} \mathbb{I}[c_i = j],\ \mathbb{I}[True] = 1, \mathbb{I}[Faulse] = 0, \quad \sum_{j=1}^{n} j s_j = n,\ \sum_{j=1}^{n} s_j = k,$$

で表す. $s = (s_1, \ldots, s_n)$ を**寸法指標** (size index) と呼ぶ.

2　ピットマン確率分割

確率分割のなかでもっとも代表的なのは**ピットマン確率分割** (Ewens-Pitman sampling formula), EPSF (θ, α) である.

$$w(n; s) := \mathbb{P}\{S_n = s\} = \frac{(\theta|-\alpha)_k}{(\theta|-1)_n} \pi_n(s) \prod_{j=1}^{n} ((1-\alpha|-1)_{j-1})^{s_j}, \tag{1}$$

$$(a|b)_k := \prod_{j=1}^{k} (a + (j-1)b),\ \pi_n(s) = \frac{n!}{\prod_{j=1}^{n} s_j!(j!)^{s_j}}, \quad \text{if } s = (s_1, \ldots, s_n) \in \mathcal{P}_{nk}.$$

パラメータ空間は

$$0 \le \alpha \le 1, -\alpha \le \theta, \quad \text{or } \alpha < 0, \theta = -M\alpha,\ M = 1, 2, \ldots$$

そのもっとも重要な性質は **partition structure**:

$$\begin{aligned} & w(n;(s_1+1,s_2,\dots,s_n))\frac{s_1+1}{n} \\ +\sum_{j=2}^{n} & w(n;(s_1,\dots,s_{j-1}-1,s_j+1,\dots,s_n))\frac{(s_j+1)j}{n}\mathbb{I}[s_{j-1}>0] \\ & = w(n-1;(s_1,\dots,s_{n-1})), \end{aligned} \tag{2}$$

である. これは n 個の個体の一つを等確率で択んで除いても確率分割法則が変わらないことを示す.

種の数　ピットマン確率分割の主要統計量である種の数 $K_n := \sum_{j=1}^{n} S_j$ の性質は良く知られている. その pmf, $\mathbb{P}\{K_n = k\} =: f_n(k)$ は前進方程式

$$f_{n+1}(k) = \frac{n-k\alpha}{\theta+n} f_n(k) + \frac{\theta+(k-1)\alpha}{\theta+n} f_n(k-1),\ 1 \le k \le n, \tag{3}$$

を満たし, 次のように陽に表せる.

$$f_n(k) = \frac{1}{(\theta|-1)_n} S_{n,k}(\theta|-1)_n(\theta|-\alpha)_k,\ 1 \le k \le n, \tag{4}$$

ただし $S_{n,k} := S_{n,k}(-1,-\alpha,0)$ で $S_{n,k}(a,b,c)$ は一般スターリング数である. この分布を EPSF-K (θ,α) で表す.

これから EPSF (θ,α) の条件付き分布

$$w(s;n,k) := \mathbb{P}\{S=s|(S\in\mathcal{P}_{n,k})\} = \frac{1}{S_{n,k}(-1,-\alpha,0)} \prod_{j=1}^{n} \frac{1}{s_j!} \left(\frac{(1-\alpha|-1)_{j-1}}{j!}\right)^{s_j}. \tag{5}$$

が定まる. これが θ に依らないことに注意.

ピットマン確率分割を拡張した**順列置換不変ギッブス確率分割** (exchangeable Gibbs random partition) の概念がある. これが上の条件により特徴付けることができる. 次節の議論は式 (5) だけに基づき, 従って順列置換不変ギッブス確率分割一般について成り立つ. ただし陽な形に表せるものは少ない.

3　種の数 $f_n(k)$ の逆過程

$f_n(k)$ は三角配列 $\{(n,k);\ 1 \le k \le n < \infty\}$ の上のマルコフ過程, あるいは酔歩と考えられる. この節ではその逆過程を考える.

$w(s;\nu,\kappa)$ からのランダム・サンプリング　条件付きピットマン確率分割 (5) から 1 個の個体をランダムに等確率で択んで除くことを考える. あるいは $\tau \in \mathcal{P}_{\nu,\kappa};\ \tau = (\tau_1,\dots,\tau_\kappa)$ の一つを除く. もし τ_1 の一つを除くと τ_1 が 1, 従って κ が 1 減少する. もし τ_j の一つを除くと τ_j が 1

減少し, τ_{j-1} が 1 増加する. κ は変化しない. これら 2 事象がそれぞれ確率 $E(S_1/\nu|S_n \in \mathcal{P}_{\nu,\kappa})$, $\sum_{j=2}^{\nu} E(jS_j/\nu|S_n \in \mathcal{P}_{\nu,\kappa})$ で生起する. 種の数を K_ν で表すと, K_ν から $K_{\nu-1}$ への推移確率は

$$\mathcal{P}\{K_{\nu-1} = \kappa - 1|K_\nu = \kappa\} = S_{\nu-1,\kappa-1}/S_{\nu,\kappa}$$
$$\mathcal{P}\{K_{\nu-1} = \kappa|K_\nu = \kappa\} = (\nu - 1 - \kappa\alpha)S_{\nu-1,\kappa}/S_{\nu,\kappa} = 1 - S_{\nu-1,\kappa-1}/S_{\nu,\kappa}.$$

この酔歩において個体数 n における種の数 K_ν の分布 (初期条件に依存する) を一般に $g_n(k) := \mathcal{P}\{K_n = k\}$ で表す. 上の推移確率から $g_n(k)$ は次の後退方程式を満たす.

$$g_n(k) = (n - k\alpha)\frac{S_{n,k}}{S_{n+1,k}}g_{n+1}(k) + \frac{S_{n,k}}{S_{n+1,k+1}}g_{n+1}(k+1), \quad 1 \le k \le n. \tag{6}$$

この式により $S|(S \in \mathcal{P}_{\nu,\kappa})$ からの確率標本の pmf, $g_n(k)$, $1 \le k \le n$ が漸次定まる. 詳しくは酔歩 $g_k(n) = g_k(n;\nu,\kappa,\alpha)$ は $g_\kappa(\nu;\nu,\kappa,\alpha) = 1$ から出発して $g_1(1;\nu,\kappa,\alpha) = 1$ に到る, 平行四辺形 $\diamond[\nu,\kappa] := \{(n,k); \max(1, n-\nu+\kappa) \le k \le \min(\kappa, n)\}$ の上で (6) を満たすが, 境界では

$$g_k(n;\nu,\kappa,\alpha) = 0, k > \min(\kappa, n), \quad \& \quad S_{n,1}/S_{n+1,1} = 1/(n-\alpha).$$

この逆過程で, 大きさ n のサブサンプルの種の期待数が次式で定まる:

$$E(K_n|S \in \mathcal{P}_{\nu,\kappa}) = \frac{1}{S_{\nu,\kappa}} \sum_{k=1}^{\min(\kappa,\nu-n)} S_{\nu-k,\kappa-1}(1-\alpha|-1)_{k-1}\left(\binom{\nu}{k} - \binom{\nu-n}{k}\right), \quad 1 < n < \nu. \tag{7}$$

(6) より三角配列上の (ν,κ) から $(\mu,\lambda) \in \diamond[\nu,\kappa]$ に到る下降酔歩が定まり, その逆の上昇酔歩が定まるがその表現は複雑となる.

4 予測

前節の結果に基づき, 次の予測を考える. 観測・捕獲データ $t \in \mathcal{P}_{n_0,k_0}$ を $S|S \in \mathcal{P}_{\nu,\kappa}$ からの確率標本と前提し, t から κ を推定する. α は t による適当な推定値を用いるとし, ν は $\nu = n_0 + 1, n_0 + 2, \ldots$ と動かすこともできるが, 任意に固定する. $(n_0,k_0) \in \diamond[\nu,\kappa]$ の条件から可能な値は $k_0 \le \kappa \le k_0 + \nu - n_0$ に限られる.

補題 1. $\kappa = \mathrm{argmax}_\lambda g_{n_0}(k_0;\nu,\lambda)$ とすると

$$\mathrm{argmax}_k g_{n_0}(k;\nu,\kappa) = k_0.$$

つまり k_0 は $g_{n_0}(k;\nu,\kappa)$ のモードである.

従って $g_{n_0}(k;\nu,\lambda)$ のモード $m(\lambda) = m(\lambda;n_0,\nu.\alpha)$ の確率を最大にする

$$\kappa_0 := \mathrm{argmax}_\lambda g_{n_0}(m(\lambda);\nu,\lambda)$$

が, 最尤推定量である. 性能の評価は数値計算によることになる.

5　他のサブサンプルと予測

EPSF (θ,α) からの単純サブサンプルは partition structure により同じ EPSF (θ,α) となり推測方法に変わりはない. EPSF-K (θ,α) で, $K_\nu=\kappa$ の条件の下での $K_n; n>\nu$ の行動を考えると

$$\mathbb{P}\{K_{\nu+n}=\kappa+\ell|K_\nu=\kappa\}=\frac{(\theta+\kappa\alpha|-\alpha)_\ell}{(\theta+\nu|-1)_n}S_{n,\ell}(-1,-\alpha,\nu-\kappa\alpha)\quad 0\le\ell\le n,$$

これは EPSF-K (θ,α) の拡張である.

$\tau\in\mathcal{P}_{\nu,\kappa}$ からの単純サブサンプルにより $\tau\in\mathcal{P}_n, n<\nu$ 上の確率分割が得られる. この確率分割から κ を予測するのに前節と同じ議論ができるが, 組合せ論的計算が障害となる.

6　補足

孤立個体数　$S|(S\in\mathcal{P}_{\nu,\kappa})$ からの大きさ n のサブサンプルにおける, 孤立個体数 S_1 を考える. その条件付き階乗モーメントは次の通りである.

$$E((S_1)_r|(S\in\mathcal{P}_{\nu,\kappa}))=(n)_r\sum_{k=1}^{n}g_n(k)\frac{S_{n-r,k-r}}{S_{n,k}},\quad 1<n<\nu,\ r=1,2,\dots \tag{8}$$

本文の (ギッブス分割の条件下での) 議論では (n,K_n) の行動だけを議論しており, これが与えられた条件の下での確率分割は常に (5) である.

参考文献

[1] Gnedin, A. and Pitman, J. (2006) Exchangeable Gibbs partitions and Stirling triangles, *Mathematical Sciences*, **138**, 5674–5685. (original —Russian version: Zapiski Nauchnnykh Seminarov ROMI, **325**, 2005, 83–102.)

[2] Lijoi, A., Prünster, I. and Walker, S.G. (2008) Baysian nonparametric estimators derived from conditional Gibbs structures, *The Annals of Applied Probability*, **18-4**, 1519–1547.

[3] Pitman, J. (2006) *Combinatorial Stochastic Processes*, Lecture Notes in Mathematics, **1875**, Springer, New York, NY.

[4] Sibuya, M. (2013) Prediction in Ewens-Pitman Sampling Formula and random samples from number-partitions, *Annals of the Institute of Statistical Mathematics*, (in print).