Quelques modèles d'attachement préférentiel

Alain Jean-Marie[1]

[1]INRIA
LIRMM CNRS/Univ. Montpellier 2

Journée ARC POPEYE
Avignon, le 9 avril 2009

Plan de l'exposé

Situation

Motivations

Modèles d'évolution de populations.

Attachement

Un processus dynamique au cours duquel des individus arrivent l'un apres l'autre et doivent choisir de rejoindre une "classe".

On cherche à expliquer les disparités de population observées : "lois d'échelle", "lois de puissance" ?
Exemples de Simon (1951) et + modernes

- villes
- mots dans un texte
- publications scientifiques, nombre de liens dans une page web
- revenus
- nombre d'espèces biologiques, allèles d'un gène
- ...

On cherche une explication endogène. Une piste :

Attachement préférentiel

La classe n'est pas choisie “au hasard” (i.e. uniformément parmi les classes) mais en fonction de la population présente

$\Longrightarrow$ les modèles d'urnes sont adaptés : contenu des urnes $\equiv$ population de la classe.

Également, modeles de graphes : individus et classes sont confondus, et ils s'attachent les uns aux autres

Statistiques recherchées : distribution des populations, valeurs limites quand beaucoup d'individus, ratio des populations dans les classes

Le modèle d'urne standard

Point de départ : un modèle d'attachement sans préférence particulière.

N urnes.

Règle

Les boules arrivent une après l'autre. Chaque boule jetée "au hasard" ($\equiv$ uniformément) dans une urne.

$\Longrightarrow$ pas de préférence entre les urnes (ou préférences constantes au cours du temps)

Évolution du nombre de boules dans les N urnes : chaîne de Markov homogène $X_n = (X_{1,n}, \ldots, X_{N,n})$

$$X_{n+1} \quad = \quad X_n \; + \; (0, \ldots, 0, 1, 0, \ldots, 0) \quad \text{avec proba } \frac{1}{K} \; .$$

$\uparrow$ $i$ième coordonnée

Comportement asymptotique

Théorème (Loi des grands nombres !)

Presque sûrement, et pour toute condition initiale,

$$\frac{1}{n} X_n \rightarrow \left(\frac{1}{N}, \ldots, \frac{1}{N}\right).$$

En particulier, les ratios

$$\frac{X_{i,n}}{X_{j,n}}$$

convergent p.s. vers des valeurs constantes, prévisibles.

Situation

Le modèle d'Eggenberger-Pólya

Une urne, contenant initialement des boules de couleur.

Règle de l'urne de Pólya-Eggenberger

Une boule est prise au hasard dans l'urne.
Elle est remise dans l'urne avec s boules de la même couleur.

$\Longrightarrow$ préférence proportionnelle à la population présente.

Cas de deux couleurs (Rouge, Bleu).

Évolution du nombre de boules dans l'urne : chaîne de Markov homogène $Z_n = (R_n, B_n)$

$$Z_{n+1} = \begin{cases} (R_n + s, B_n) & \text{avec proba } \dfrac{R_n}{R_n + B_n} \\ \\ (R_n, B_n + s) & \text{avec proba } \dfrac{B_n}{R_n + B_n} \end{cases} .$$

Propriétés

s objets de plus à chaque étape : $S_n \equiv R_n + B_n = R_0 + B_0 + ns$.
Soient les ratios :

$$\rho_n = \frac{R_n}{R_n + B_n} \qquad \beta_n = \frac{B_n}{R_n + B_n} .$$

Théorème

Les suites de variables aléatoires $\{\rho_n\}$ et $\{\beta_n\}$ sont des martingales par rapport à la filtration engendrée par les $\{Z_n\}$.

Preuve

$$\begin{aligned} \mathbb{E}(\rho_{n+1}|(R_n, B_n)) &= \frac{R_n + s}{S_n + s}\frac{R_n}{S_n} + \frac{R_n}{S_n + s}\frac{B_n}{S_n} \\ &= \frac{R_n(S_n + s)}{(S_n + s)(S_n)} = \rho_n . \end{aligned}$$

Loi limite

Théorème (Pólya, 1931)

Presque sûrement,

$$\frac{1}{n}\,(R_n, B_n) \quad \rightarrow \quad (U, 1-U)$$

où U a une distribution Beta(R_0, B_0).

Pour mémoire : la densité de la loi Beta(α, β) est

$$f_{\alpha,\beta}(x) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} \;.$$

En particulier, si $R_0 = B_0 = 1$, la distribution est uniforme.

Les problèmes avec ce résultat

Comme observé par Gaujal, Thierry (2007), ce résultat pose des problèmes potentiels pour les applications :

- la limite n'est pas déterministe
- elle dépend fortement de la condition initiale
- le ratio

$$\tau_n = \frac{\max\{B_n, R_n\}}{\min\{B_n, R_n\}} \sim \frac{\max\{U, 1-U\}}{\min\{U, 1-U\}}$$

  ne converge que vers une v.a.

$\Longrightarrow$ valeur prédictive/explicative du modèle sur un petit nombre d'observations ?
Par contre, des limites p.s. peuvent apparaître pour des généralisations.

Situation

Le modèle de Simon

Une suite d'objets de différents "types" arrive. On suppose que :

- la proba que l'objet soit d'un type pas encore apparu est α
- sinon, la proba que l'objet soit d'un type k déja apparu est proportionnel à la fréquence empirique de ce type.

Dans le langage des urnes, on obtient ce résultat avec :

Règles de l'urne de Simon

Une boule est prise au hasard dans l'urne.
Elle est remise dans l'urne avec une seconde boule :

- d'une nouvelle couleur avec proba α
- de la même couleur avec proba $1 - \alpha$

Se réduit au modèle d'Eggenberger-Pólya si $\alpha = 0$. Facilement analysable aussi si $\alpha = 1$...

L'analyse de Simon

Soit $f(k,K)$ le nombre moyen d'urnes avec k boules à l'étape K.

$$f(k,K+1) \;=\; f(k,K) + \frac{1-\alpha}{K}\left((k-1)f(k-1,K) - kf(k,K)\right) \;.$$

De cette récurrence on "déduit" que

$$\frac{1}{K}f(k,K) \;=\; \frac{\alpha}{2-\alpha}\frac{\Gamma(k)\Gamma(2+1/\bar{\alpha})}{\Gamma(k+1+1/\bar{\alpha})} \;\sim\; k^{-1-1/\bar{\alpha}} \;.$$

Mais on a le problème qu'il ne s'agit que d'une moyenne.

Solution exacte

Kullmann et Kertész (2008) ont trouvé la distribution.
Pour $X_0 = (1)$ et avec $\bar{\alpha} = 1 - \alpha$,

$$\mathbb{P}(X_{i,K} = k) = \alpha^{i-1} \sum_{\ell=1}^{k} (-1)^{\ell-1} \binom{k-1}{\ell-1} \frac{\Gamma(K - \ell\bar{\alpha})}{\Gamma(k)\Gamma(1-\ell\bar{\alpha})} \left[\sum_{b=i}^{K} \frac{\Gamma(b)\Gamma(1-\ell\bar{\alpha})}{\Gamma(b-\ell\bar{\alpha})} \binom{b-2}{i-2} \bar{\alpha}^{b-i} \right].$$

Par exemple, la "distribution moyenne de la taille"

$$\lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \mathbb{P}(X_{i,K} = k) = \frac{\Gamma(k)\Gamma(2+1/\bar{\alpha})}{\Gamma(k+1+1/\bar{\alpha})} \frac{\alpha}{1+\bar{\alpha}} \sim k^{-1-1/\bar{\alpha}}.$$

Situation

L'urne de Hoppe

Une seule urne, une couleur spéciale (Noir) et une infinité de couleurs non-noires. Initialement, β boules noires.

Règle de l'Urne de Hoppe

On tire une boule au hasard dans l'urne. On la remet avec une seconde boule :

- d'une nouvelle couleur si la boule tirée est noire
- de la même couleur sinon.

On crée donc une nouvelle couleur avec probabilité

$$\frac{\beta}{\beta + \sum_i n_i} .$$

Modèle introduit dans Hoppe (1984), analysé dans Hoppe (1987).

Si $\beta = 0$, c'est le modèle de Eggenberger-Pólya.
Si $\beta = 1$, c'est le modèle du Restaurant Chinois.

Nouvelles couleurs

Distribution du nombre N_K de nouvelles couleurs

Fonction génératrice, moyenne... (Ewens 1972)

$$
\begin{aligned}
\mathbb{E}\left(u^{N_K}\right) &= \prod_{\ell=0}^{K-1} \frac{\ell+\beta u}{\ell+\beta} \\
\mathbb{E}N_K &= \sum_{\ell=0}^{K-1} \frac{\beta}{\ell+\beta} \sim \beta \log K \\
\mathbb{V}N_K &\sim \beta \log K \ .
\end{aligned}
$$

Comportement asymptotique :

Théorème (Mahmoud)

$$
\frac{N_K - \beta \log K}{\sqrt{\log K}} \to_{\mathcal{L}} \mathcal{N}(0, \beta) \ .
$$

Distribution d'occupation

Soit A_ℓ le nombre d'urnes dont la population est ℓ.

Loi d'occupation (Ewens, Karlin & McGregor, Hoppe)

Pour toute partition $\mathbf{a} = (a_1, \dots, a_K)$ du nombre K ($K = \sum i a_i$), on a :

$$\mathbb{P}(\mathbf{A} = \mathbf{a}) \; = \; K! \prod_{\ell=1}^{K} \frac{\beta^{a_\ell}}{a_\ell! \ell^{a_\ell} (\beta + \ell - 1)} \, ,$$

dite "Ewens' sampling formula".

Situation

Une approche différente

Introduction d'un nouveau modèle.
Raisons :

- pour pouvoir calculer la distribution des populations,
- pour avoir une croissance moins que linéaire du nombre de couleurs

Une approche différente

Introduction d'un nouveau modèle.
Raisons :

- pour pouvoir calculer la distribution des populations, ne sachant pas que c'est possible pour le modèle de Simon
- pour avoir une croissance moins que linéaire du nombre de couleurs, ne connaissant pas le modèle de Hoppe.

Une approche différente

Introduction d'un nouveau modèle.
Raisons :

- pour pouvoir calculer la distribution des populations, ne sachant pas que c'est possible pour le modèle de Simon
- pour avoir une croissance moins que linéaire du nombre de couleurs, ne connaissant pas le modèle de Hoppe.

Constatation :

- le modèle de Pólya-Eggenberger est relativement facile à analyser grâce au fait que les ajouts de boules "commutent en probabilité".
- dans le modèle de Simon qui le généralise, on a perdu cette commutation.

$\Longrightarrow$ rétablir la situation !

Commutativité

Deux événements quand la population est $n = |\mathbf{n}|$:

- arrivée dans la couleur i : $A_i(\cdot)$, proba $(1-\alpha(n))n_i/n$
- nouvelle couleur : $N(\cdot)$, proba $\alpha(n)$

Probabilité de $A_i(N(\cdot))$:

$$\alpha(n)\ (1-\alpha(n+1))\ \frac{n_i}{n+1}\ .$$

Probabilité de $N(A_i(\cdot))$:

$$(1-\alpha(n))\ \frac{n_i}{n}\alpha(n+1)\ .$$

Égalité si :

$$\alpha(n) \ = \ \frac{\beta}{n+\beta} \qquad \beta \equiv \frac{\alpha_1}{1-\alpha_1}\ .$$

Règles de l'urne

Règle de l'urne " commutative "

On tire une boule au hasard dans l'urne. On la remet avec :

- une boule d'une nouvelle couleur avec probabilité $\alpha(n)$
- une boule de la même couleur avec probabilité $(1 - \alpha(n))$

où

$$\alpha(n) \;=\; \frac{\beta}{n+\beta}\;.$$

Et c'est l'Urne de Hoppe avec β boules noires ! !

Comme dans ce modèle, toutes les évolutions ont la même proba. Il suffit de dénombrer les chemins d'un état à un autre.

Loi de transition

Soient **n** et **m** deux vecteurs :

- $\mathbf{n} \in \mathbb{S}_d \equiv \{(n_1, \ldots, n_d),\ n_i \in \mathbb{N}^*, 1 \leq i \leq d\}$,
- $\mathbf{m} \in \mathbb{S}^+_{d,d'} \equiv \{(n_1, \ldots, n_{d'}),\ n_i \in \mathbb{N}, 1 \leq i \leq d$, $n_i \in \mathbb{N}^*, d+1 \leq i \leq d'\}$.

Alors :

Probas de transition

$$\mathbb{P}(\mathbf{n} \to \mathbf{n} + \mathbf{m})$$
$$= \frac{\beta^{d'-d}}{\prod_{j=|\mathbf{n}|}^{|\mathbf{n}|+|\mathbf{m}|-1} j + \beta} \frac{|\mathbf{m}|!}{(d'-d)!} \prod_{i=1}^{d} \binom{n_i + m_i - 1}{n_i - 1} \prod_{j=d+1}^{d'} \frac{1}{m_j} .$$

Quelques identités

Identité plutôt connue :

$$\sum_{\substack{\mathbf{m} \in \mathbb{N}^d \\ |\mathbf{m}|=K}} \prod_{i=1}^{d} \binom{n_i + m_i - 1}{n_i - 1} = [z^K] \frac{1}{(1-z)^{|\mathbf{n}|}} = \binom{|n| + K - 1}{K} .$$

Identité moins connue (?) :

$$\sum_{\mathbf{m} \in (\mathbb{N}^*)^d, |\mathbf{m}|=K} \prod_{i=1}^{d} \frac{1}{m_i} = [z^K] \left(\log \frac{1}{(1-z)} \right)^d = ?$$

Fonctions génératrices

Un vecteur $\mathbf{n} \in \mathbb{S}_d$ étant donné :

La FG maîtresse

$$\sum_{k=0}^{\infty} u^k \sum_{\mathbf{m} \in \mathbb{S}^+_{d,d+k}, |\mathbf{m}|=K} \prod_{i=1}^{d+k} z_i^{m_i} \, \mathbb{P}(\mathbf{n} \to \mathbf{n}+\mathbf{m})$$

$$= \frac{\Gamma(|\mathbf{n}|+\beta)K!}{\Gamma(|\mathbf{n}|+\beta+K)} [z^K] \prod_{i=1}^{d} \frac{1}{(1-zz_i)^{n_i}}$$

$$\sum_{k=0}^{\infty} \frac{(\beta u)^k}{k!} \prod_{j=d+1}^{d+k} \log \frac{1}{(1-zz_j)} .$$

Nouvelles couleurs

Distribution du nombre N_K de nouvelles couleurs

Nouvelles couleurs

$$
\begin{aligned}
\mathbb{E}\left(u^{N_K}\right) &= \prod_{\ell=0}^{K-1} \frac{|\mathbf{n}| + \ell + \beta u}{|\mathbf{n}| + \ell + \beta} \\
\mathbb{E} N_K &= \sum_{\ell=0}^{K-1} \frac{\beta}{|\mathbf{n}| + \ell + \beta} \sim \beta \log K \\
\mathbb{V} N_K &= \sum_{\ell=0}^{K-1} \frac{\beta}{|\mathbf{n}| + \ell + \beta} \frac{|\mathbf{n}| + \ell}{|\mathbf{n}| + \ell + \beta} \sim \beta \log K \ .
\end{aligned}
$$

Ce sont les résultats de Ewens/Hoppe modifés. Voir Pitman.

Répartition dans les couleurs

FG des populations des "vieilles couleurs" et de la population émigrante E_K :

FG nouveau monde/ ancien monde

$$\mathbb{E}\left(\prod_{i=1}^{d} z_i^{X_{i,K}} z_0^{E_K}\right)$$

$$= \frac{\Gamma(|\mathbf{n}|+\beta)K!}{\Gamma(|\mathbf{n}|+\beta+K)}[z^K]\prod_{i=1}^{d}\frac{1}{(1-zz_i)^{n_i}}\frac{1}{(1-zz_0)^{\beta}}.$$

Valeurs

Nombres moyens dans les anciennes couleurs i

$$\mathbb{E}X_{i,K} = \frac{Kn_i}{|\mathbf{n}| + \beta} \qquad \mathbb{E}E_K = \frac{K\beta}{|\mathbf{n}| + \beta}.$$

Variances...

Pour la répartition dans les nouvelles couleurs...

Nombre moyen dans la nouvelle couleur j

$$\mathbb{E}X_{j,K} = \frac{\Gamma(|\mathbf{n}| + \beta)K!}{\Gamma(|\mathbf{n}| + \beta + K)} [z^K] \frac{z}{(1-z)^{|\mathbf{n}|+1}} \sum_{k=j}^{\infty} \frac{\beta^k}{k!} \left(\log \frac{1}{1-z}\right)^{k-1}.$$

Propriétés asymptotiques

Lois limites dans les anciennes couleurs : voir Eggenberger-Pólya.

Nombre de nouvelles couleurs : voir Hoppe.

$$\frac{1}{\log K} N_K \ \to_{\mathcal{L}} \ \beta \ .$$

Population des nouvelles couleurs : pour tout $j = O(1)$,

$$\mathbb{E} X_{j,K} \ \sim \ \frac{1}{|\mathbf{n}| + \beta} \frac{\log K}{K} \ .$$

Convergence pas bien claire pour $K = O(100)$...

Fréquences et occupations limites

Proportions limites (Pitman)

Le vecteur de fréquences de population $(X_{j,K}/K; j = 1, \ldots)$ converge vers le vecteur aléatoire

$$(W_1, (1 - W_1)W_2, (1 - W_1)(1 - W_2)W_3, \ldots)$$

où les W_i sont indépendantes et $\sim \text{Beta}(1, \beta)$.

Occupations limites (Pitman)

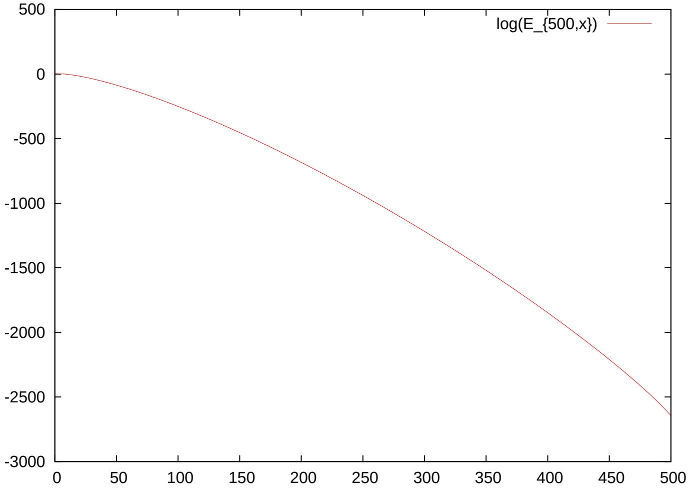
Le vecteur de fréquences d'occupation $(\#\{j|X_{j,K} = \ell\}, \ell = 1, \ldots)$ converge vers le vecteur aléatoire

$$(Z_{\beta,1}, Z_{\beta,2}, \ldots)$$

où les $Z_{\beta,i}$ sont indépendantes et $\sim \text{Poisson}(\beta/i)$.

Repartition des populations

En moyenne, les nouvelles couleurs sont très inégales...



Généralisation de Pitman

Pitman a introduit un deuxième paramètre.

Règle de Pitman

Les boules arrivent une par une. Sachant qu'il y a k couleurs, la boule n est :

$$\begin{array}{ll} \text{est d'une nouvelle couleur} & \text{avec probabilité } \dfrac{\beta + k\alpha}{n+\beta}\,, \\ \text{de la couleur } i & \text{avec probabilité } \dfrac{n_i - \alpha}{n+\beta}\,. \end{array}$$

Ce modèle est également "commutatif".

$\Longrightarrow$ certains résultats se généralisent.

Des asymptotiques en n^α apparaissent pour $0 < \alpha < 1$.

Perspectives

Parmi les questions en suspens :

- loi limite pour les populations moyennes ?
- loi limite pour les statistiques d'ordre ?
- extension à des urnes “multicolores” plus générales...

Bibliographie

Modèles d'Urnes

N.J. Johnson and S. Kotz, *Urn Models and Their Application*, Wiley, 1977.

H.M. Mahmoud, *Pólya Urn Models*, CRC Press, 2009.

J. Pitman, "Combinatorial Stochastic Processes", Springer, 2006.

F.M. Hoppe, "Pólya-like urns and the Ewens' sampling formula", *J. Math. Biology*, 20, pp. 91–94, 1984.

F.M. Hoppe, "The sampling theory of neutral alleles and an urn model", *J. Math. Biology*, 25, pp. 123–159, 1987.

Bibliographie (suite)

Attachement préférentiel

- L. Kullmann and J. Kertész, "Preferential growth : exact solution of the time dependent distributions", `arXiv :cond-mat/0012410v1`, 2008.

Modèles de population

- H.A. Simon, "On a class of skew distribution functions", *Biometrika*, 42, 3/4, pp. 425–440, 1951.
- B. Gaujal, L. Gulyas, Y. Surdati Mansuri and E. Thierry, "Markov chain analysis of an agent growth model", LIP Research Report, 2007-15, 2007.