Inférence statistique sur des processus de branchements multi-types

Amel Ouaari [1] & Rachid Senoussi [2]

[1] *Biostatistique et Processus Spatiaux (BioSP), INRA, Site Agroparc,84914 Avignon, France. et amel.ouaari@paca.inra.fr*

[2] *Biostatistique et Processus Spatiaux (BioSP), INRA, Site Agroparc,84914 Avignon, France. et rachid.senoussi@paca.inra.fr*

Résumé.
On s'intéresse aux estimateurs du maximum de vraisemblance des paramètres associés à des processus de branchements en temps continu, multi-types et homogènes. L'exposé passe par la description probabiliste et paramétrée de l'évolution temporelle de plusieurs populations ou classes en interaction. Nous décrivons le comportement du système multi-types via l'évolution des fonctions génératrices de la loi des effectifs et faisons le lien avec la théorie des équations différentielles aux dérivées partielles. Notre approche statistique consiste à inverser théoriquement ou numériquement la fonction génératrice en utilisant le théorème de Cauchy pour les fondrées anaytiques pour inférer ces paramètres.

Mots-clés. Processus de branchements en temps continu multi-types, estimateurs de maximum de vraisemblance, équations aux dérivées partielles, théorème de Cauchy.

Abstract.
This paper deals with the maximum-likelihood estimators of parameters associated to homogeneous multi-type continuous-time branching processes. It describes, in a probabilistic way, the temporal change of sever interacting populations or classes. We first describe the behaviour of the multi-type system through the evolution of the generating functions of the offspring distribution and link it to Partial Differential Equation theory. Our statistical approach consists in inverting theoretically or numerically the probability-generating functions using Cauchy's theorem in order to infer these parameters.

Keywords. Multi-type continuous-time branching processes, maximum-likelihood estimators, Partial Differential Equations, Cauchy's theorem.

1 Introduction

L'intérêt des processus de branchement multi-types réside dans les nombreuses applications dans des domaines aussi divers que la biologie, l'épidémiologie, la démographie, la physique des particules, etc. Les processus de branchements sont considérés comme des modèles de probabilité appropriés pour la description du comportement des systèmes dont

les composantes (cellules, particules, $\cdots$ etc) se reproduisent, se transforment, interagissent ou meurent. Par ailleurs, en dynamique de populations, il y a un manque flagrant de lois de probabilité modélisant des comptages simultanés des individus de plusieurs populations interactives et évoluant dans le temps. L'intérêt de développer des méthodes d'inférence statistique propres à ces modèles est de grande actualité. Il existe de nombreuses approches d'estimation paramétriques dans les processus de branchements, mais celles-ci sont pour la plupart basées sur des méthodes de contrastes et de propriétés de martingale [6]. Nous proposons ici de revenir à la méthode du maximum de vraisemblance en inversant les équations des fonctions génératrices.

Notations:

Nous utilisons les notations suivantes:

K sera le nombre de populations en jeu, et $\boldsymbol{i} = (i_1, \cdots, i_K)$, multi-indice de $\mathbb{N}^K$, décrira les tailles respectives des différents sous populations.

De même, $\mathbf{0} = (0, \cdots, 0)$, $\mathbf{1} = (1, \cdots, 1)$ *et* $\boldsymbol{e}_i = (0, \cdots, , 0, \overbrace{1}^{i}, 0, \cdots, 0)$: seront des vecteurs de $\mathbb{N}^K$ d'utilisation constante.

On notera $\mathbf{X(t)} = (X_1(t), \cdots, X_K(t)) \in \mathbb{N}^K$ les tailles des populations à l'instant t et si $\boldsymbol{z} = (z_1, \cdots, z_K)$ est dans $\mathbb{C}^K$, on écrira $\boldsymbol{z}^{\mathbf{i}} = z_1^{i_1} \times \cdots \times z_K^{i_K}$

2 Equations du branchement multi-types

On considère les K fonctions génératrices de $X(t)$, quand on part avec un état initial $X(0) = \boldsymbol{e}_i, \quad i = 1, \cdots, K$

$$G_i(t, \boldsymbol{z}) = E(\boldsymbol{z}^{\mathbf{X(t)}} / X(0) = \boldsymbol{e}_i).$$

Si on ne considère que des branchements homogènes dans le temps, alors la notation vectorielle: $G(t, \boldsymbol{z}) = (G_1(t, \boldsymbol{z}), \cdots, G_K(t, \boldsymbol{z}))$ permet de traduire simplement l'équation de Chapman-Kolmogorov du branchement multi-type par:

$$G(t + s, \boldsymbol{z}) = G(t, G(s, \boldsymbol{z})), \quad \textit{pour} \quad \textit{tous} \quad s, t \geq 0 \quad \textit{et} \quad \boldsymbol{z} \in D(0, 1)^K \subset \mathbb{C}^K$$

avec la condition initiale $G(0, \boldsymbol{z}) = \boldsymbol{z}$.

Remarques:

- Si $X(0) = \boldsymbol{m} = (m_1, \cdots, m_K)$ alors $G_{\boldsymbol{m}}(t, \boldsymbol{z}) = G_1^{m_1}(t, \boldsymbol{z}) \times \cdots \times G_K^{m_K}(t, \boldsymbol{z})$.
- Si $X(0)$ est de loi γ_0 sur $\mathbb{N}^K$ et de génératrice G_0, alors: $G(t, \boldsymbol{z}) = G_0(G(t, \boldsymbol{z}))$.

$X(t)$ est un processus markovien de sauts, homogène dans le temps. On peut de manière explicite exprimer le générateur infinitésimal Q du semi groupe

$$P_{i,j}(t) = P(X(t) = j \setminus X(0) = i)$$

$$Q_{i,j} = \lim_{t \longrightarrow 0} \frac{P_{i,j}(t) - \mathbb{I}_{\{i=j\}}}{t}$$

On a alors pour $\boldsymbol{z} \in D(0,1)^K$ et pour $V_j(\boldsymbol{z}) = \partial_t G_j(0, \boldsymbol{z})$, l'égalité suivante:

$$V_j(\boldsymbol{z}) = Q_{j,j}\boldsymbol{z}^j + \sum_{l \neq j} Q_{j,l}\boldsymbol{z}^l.$$

En écriture vectorielle, on en déduit pour $V(\boldsymbol{z}) = \partial_t G(0, \boldsymbol{z})$, que

$$\partial_t G(t, \boldsymbol{z}) = -V(\boldsymbol{z})(\partial_{\boldsymbol{z}} \mathbf{G}(t, \boldsymbol{z}))^T. \tag{1}$$

Notons que, chaque colonne de l'équation dynamique (en t) ne met en jeu qu'un seul G_i à la fois (et les dérivées en $t = 0$ des autres G_i).
Il s'agit donc de K équations aux dérivées partielles, linéaires homogènes, identiques. dont, seules les conditions initiales varient.
Fonction génératrice des moments
Rappelons que si X est une v.a sur $\mathbb{N}^K$ de loi $P = (p_{\boldsymbol{m}},\ \boldsymbol{m} \in \mathbb{N}^K)$ a pour génératrice

$$G(\boldsymbol{z}) = E(\boldsymbol{z}^{\mathbf{X}(t)}) = \sum p_{\boldsymbol{m}} \boldsymbol{z}^{\boldsymbol{m}}$$

alors, G est (multi) analytique en $\boldsymbol{z}$ dans $(D(0,1))^K$, et que les dérivés partielles d'ordre m en 0, redonnent la loi de probabilité P :

$$\partial^{m_1 \cdots m_K}_{z_1^{m_1} \cdots z_1^{m_1}} G(0) = m_1! \cdots m_K! p_{\boldsymbol{m}}.$$

3 Approche statistique par Max-Vraisemblance

Pour l'interprétation du modèle de branchement, la paramétrisation des lois de transition $P_{i,j}(t)$ ou de façon équivalente des génératrices $G_i(t, \boldsymbol{z})$ se fera naturellement par le biais du générateur infinitésimal $Q_{i,j}$, autrement dit des fonctions analytiques $V_j(\boldsymbol{z})$.

- $-Q_{j,j}$: est le paramètre de la loi exponentielle du temps de vie d'un individu de la population j.

- $\frac{Q_{j,m}}{-Q_{j,j}}$: est la probabilité pour qu'un individu de la population j produise $\boldsymbol{m} = (m_1, \cdots, m_K)$ descendants lors de son décès .

Le problème d'inférence réside donc dans l'inversion des fonctions génératrices après avoir résolu les équations dynamiques (1) afin d'extraire les probabilités correspondantes.

3.1 Vraisemblance avec des observations ponctuelles de $X(t)$ dans le temps $t_1, \cdots, t_m$

La propriété de Markov dit que:

$$\mathcal{L}(X(t_1), \cdots, X(t_m)/X(0)) = \mathcal{L}(X(t_1)/X(0)) \cdots \mathcal{L}(X(t_m)/X(t_{m-1}))$$

et l'hypothèse d'homogénéité nous dit aussi que pour les observations dans $\mathbb{N}^K$

$$X(0) = n_0,\ X(t_1) = n_1, \cdots, X(t_m) = n_m.$$

La vraisemblance s'écrit :

$$V_\theta = P_{n_0,n_1}(t_1 \mid \theta) P_{n_1,n_2}(t_2 - t_1 \mid \theta) \cdots P_{n_{m-1},n_m}(t_m - t_{m-1} \mid \theta)$$

Lorsque $\theta = (\theta_1, \cdots, \theta_r)$ désigne le vecteur des paramètres d'intérêt de dimension finie r spécifiant le générateur infinitésimal.
L'estimateur du maximum de vraisemblance s'obtient par la résolution des r équations:

$$\partial_\theta log V = \sum_{i=1}^{m} \frac{\partial_{\theta_l} P_{n_{i-1},n_i}(t_i - t_{i-1} \mid \theta)}{P_{n_{i-1},n_i}(t_i - t_{i-1} \mid \theta)} = 0.$$

Ce qui nous amène à calculer dans un premier temps les fonctions génératrices des transitions $P_{n,\cdot}(s \mid \theta)$ associées à:

$$G^n(s, \boldsymbol{z} \mid \theta) = G_1^{n_1}(s, \boldsymbol{z} \mid \theta) \times \cdots \times G_K^{n_K}(s, \boldsymbol{z} \mid \theta)$$

c'est-à-dire

$$P_{n,m}(s \mid \theta) = \frac{1}{m!} \partial^{m_1+\cdots+m_K}_{z_1^{m_1} \cdots z_K^{m_K}} \mathbf{G}^{\mathbf{n}}(s, \boldsymbol{z} \mid \theta)$$

3.2 Inversion de la génératrice

Si G est une fonction analytique, pour tout $\boldsymbol{z} \in D(0,r)^K$, l'utilisation du théorème de Cauchy pour les fonctions analytiques (uni ou multivariables) permet de calculer les probabilités correspondantes

$$P_{n_1 \cdots n_K} = \frac{1}{(2\pi i)^K} \int_{\gamma_1} \int \int_{\gamma_K} \frac{G(\xi_1, \cdots, \xi_K)}{\xi_1^{n_1+1}, \cdots, \xi_K^{n_K+1}} d\xi_1 \cdots \xi_K$$

L'approximation sera d'autant meilleure quand K n'est pas très grand et les chemins γ_i simples.

3.3 Application

Partant de l'équation aux dérivées partielles:

$$\begin{cases} \frac{\partial G_i(t,\boldsymbol{z})}{\partial t} & = & \sum\limits_{j=1}^{k} -V_j(\boldsymbol{z})\frac{\partial G_i(t,\boldsymbol{z})}{\partial z_j}, \\ G_i(0,\boldsymbol{z}) & = & z_i. \end{cases}$$

On normalisera d'abord les écritures des équations car pour tout Q^* vérifiant

$$Q^* \geq sup(\overline{Q}_{jj}) \qquad o\grave{u} \qquad \overline{Q}_{jj} = -\sum_{\substack{r=0 \\ r\neq j}}^{\infty} Q_{jr}, \qquad (j = 1, \cdots, K),$$

et pour tout j, il existe $Q^{**}_{j,j}$ tel que $\overline{Q}_{jj} + Q^{**}_{jj} = Q^*$ et alors:
$V_j(\boldsymbol{z}) = Q^*[H_j(\boldsymbol{z}) - z_j]$ avec $H_j(\boldsymbol{z})$ désignant la fonction génératrice du branchement de type j, mieux interprétable.

Exemple de branchements Poissonniens indépendants

$$H_j(\boldsymbol{z}) = \prod_{i=1}^{K} e^{\lambda_{ji}(z_j-1)} = e^{\sum\limits_{i=1}^{K} \lambda_{ji}z_j} e^{\sum\limits_{i=1}^{K} \lambda_{ji}} = e^{<\vec{\lambda_j},z>} e^{-\overline{\lambda_j}}$$

où $\overline{\lambda_j} = \sum\limits_{i=1}^{K} \lambda_{ji}$

On standardisera l'EDP de la façon suivante:
Soit $\Lambda = (\lambda_{ij})$ la matrice des paramètres,
et $\Lambda^{-1} = (\mu_{ij})$, $\gamma_j = e^{-\overline{\lambda_j}}$, $\beta_{jl} = \sum\limits_{j=1}^{k}\sum\limits_{l=1}^{k} \lambda_{jl}\mu_{jl}$

alors $H_j(\Lambda\boldsymbol{u}) = e^{u_j - \overline{\lambda_j}}$

Le changement de variables $\boldsymbol{z} = \Lambda\boldsymbol{u}$ donne la fonction $\tilde{G}(t,\boldsymbol{u}) = G(t,\Lambda\boldsymbol{u})$.
Elle vérifie: $\partial_t\tilde{G}(t,\boldsymbol{u}) = \theta\sum\limits_{l=1}^{k} V_j(\boldsymbol{u})\partial_{\boldsymbol{u}_l}\tilde{G}(t,\boldsymbol{u})$
avec

$$V_j(\boldsymbol{u}) = \gamma_{j1}e^{u_1} + \cdots + \gamma_{jj}e^{u_K} - (\beta_{j1}u_1 + \cdots + \beta_{jK}u_K).$$

La résolution du système d'équations différentielles ordinaires de dimension K:

$$\begin{cases} u_1' = V_1(u_1, \cdots, u_K) \\ \vdots \\ u_K' = V_K(u_1, \cdots, u_K) \end{cases}$$

s'écrit $\boldsymbol{u'} = \gamma e^{\boldsymbol{u}} - \beta \boldsymbol{u}$
avec $\boldsymbol{u} = (u_1, \cdots u_k)^T, \quad e^{\boldsymbol{u}} = (e^{u_1}, \cdots, e^{u_k}), \quad \beta = \beta_{jl}$
Cas d'une population unique
Pour résoudre l'EDP (1), il faut alors résoudre l'équation différentielle ordinaire:

$$\frac{du}{dt} = \gamma e^u - \beta u$$

c'est-à-dire chercher une primitive F de $f(u) = \frac{1}{\gamma e^u - \beta u}$ et une intégrale première de ce système.
La génératrice solution sera la valeur z_0 satisfaisant l'équation implicite

$$F(z) - F(z_0) = t$$

Nous conclurons la présentation par des calculs numériques et des simulations illustrant la portée de cette approche d'estimation.

Bibliographie

[1] Athreya . P. E. Ney (1972), Branching Processes, Springer, Berlin.

[2] Athreya, K. B. et Niels Keiding(1977), Estimation theory for continuous-time branching processes,Indian Institute of Science, Bangalore, India, University of Copenhagen, Denmar

[3] González, C. et Minuesa, I. del Puerto(2004), Maximum likelihood estimation and expectation–maximization algorithm for controlled branching processes, Data Analysis, Badajoz, Spain.

[4] González, M. et Martín,J. et Martínez,R. et Mota, M. (2007), Non-parametric Bayesian estimation For multitype branching processes through simulation-based methods, Computaional Statistics Data Analysis, Badajoz, Spain.

[5] Neils Becker (1977), Estimation for Discrete Time Branching Precesses with Application to Epidemics , Biometrics, La Trobe university, Bundoor, Australia.

[6] Soren Asemussem et Neils Keiding (1978), Martingale central limit theorems and asymptotic estimation theory for multitype branching processes, University of Copen.