Estimation paramétrique des chaînes semi-markoviennes pour des données censurées

Vlad Stefan Barbu [1] & Caroline Bérard [2] & Dominique Cellier [3] & Mathilde Sautreuil [4] & Nicolas Vergne [5]

[1] *LMRS UMR 6085, Normandie Université, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France ; barbu@univ-rouen.fr*
[2] *LITIS EA 4108, Normandie Université, Rouen, France ; caroline.berard@univ-rouen.fr*
[3] *associé à LITIS EA 4108, Normandie Université, Rouen, France ; dominique.cellier@laposte.net*
[4] *LMRS UMR 6085, Normandie Université, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France ; mathilde.sautreuil@etu.univ-rouen.fr*
[5] *LMRS UMR 6085, Normandie Université, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France ; nicolas.vergne@univ-rouen.fr*

Résumé. L'objectif de notre présentation est double. D'une part, nous considérons le problème d'estimation paramétrique d'une chaîne semi-markovienne (cf. Barbu et. al [1]), en prenant en compte plusieurs cas : censure au début et/ou à la fin, pas de censure, une ou plusieurs trajectoires. Nous nous intéressons au cas général de noyau semi-markovien, mais aussi à des cas particuliers, importants en pratique. D'autre part, nous présentons un package `R` que nous avons développé (cf. Barbu et. al [2]). Il faut noter que dans ce package nous avons aussi implémenté une approche non-paramétrique.

Mots-clés. Estimation paramétrique, chaînes semi-markoviennes, données censurées, package `R`

Abstract. The purpose of our presentation is double. On the one hand, we consider the problem of parametric estimation of a semi-Markov chain (cf. Barbu et. al [1]), taking into account several cases : censoring at the beginning and/or at the end, no censoring, one or several trajectories. We are interested in the general semi-Markov kernel, but also in particular cases, important from practical point of view. On the other hand, we present an `R` package that we have developed (cf. Barbu et. al [2]). Note that in this package we implement also a nonparametric estimation approach.

Keywords. Parametric estimation, semi-Markov chains, censored data, `R` package

1 Chaînes semi-markoviennes

Considérons un système aléatoire avec un nombre fini d'états $E = \{1, \ldots, s\}$, $s < \infty$. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et supposons que l'évolution au cours du temps du

système est gouvernée par un processus stochastique $(Y_k)_{k\in\mathbb{N}}$, défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans E. Soient $(T_m)_{m\in\mathbb{N}}$, définis sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{N}$, les instants successifs de changement d'état dans $(Y_k)_{k\in\mathbb{N}}$ et considérons aussi $(J_m)_{m\in\mathbb{N}}$, définis sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans E, les états successivement visités à ces instants de temps. Notons aussi $X_m = T_m - T_{m-1}$, $m \in \mathbb{N}^*$, et, par convention, mettons $X_0 = T_0 = 0$. La relation entre le processus $(Y_k)_{k\in\mathbb{N}}$ et le processus $(J_n)_{\in\mathbb{N}}$ est donnée par $Y_k = J_{N(k)}$ ou $J_m = Y_{T_m}, m, k \in \mathbb{N}$, où

$$N(k) := \max\{m \in \mathbb{N} \mid T_m \leq k\} \tag{1}$$

est le processus de comptage du nombre de sauts dans $[1, k] \subset \mathbb{N}$.

Les définitions suivantes introduisent les notions de noyau semi-markovien, la notion de chaîne de renouvellement markovien (CRM) et celle de chaîne semi-markovienne (CSM).

Définition 1. *Une fonction matricielle* $\mathbf{q} := \big(q(k); k \in \mathbb{N}\big)$ *est appelée noyau semi-markovien si :*

1. $q_{ij}(k) \geq 0,\ i, j \in E,\ k \in \mathbb{N}$,
2. $\sum_{k=0}^{\infty} \sum_{j\in E} q_{ij}(k) = 1,\ i \in E$.

Définition 2. *Sous les notations précédentes :*

1. *La chaîne (J, S) est appelée chaîne de renouvellement markovien (CRM) associée au noyau q si, pour tout $k, l \in \mathbb{N}$, $i, j \in E$, vérifie $\mathbb{P}$-p.s.*

$$\begin{aligned} &\mathbb{P}(J_{m+1} = j, T_{m+1} - T_m = k \mid J_m = i, J_{m-1}, \ldots, J_0, T_m, \ldots, T_0) \\ &= \mathbb{P}(J_{m+1} = j, T_{m+1} - T_m = k \mid J_m = i) =: q_{ij}(k). \end{aligned} \tag{2}$$

2. *La chaîne $Y = (Y_n)_{n\in\mathbb{N}}$ à valeurs dans E est dite chaîne semi-markovienne (CSM) associée à (J, T).*

Notons que si (J, T) est une CRM, alors $J = (J_m)_{m\in\mathbb{N}}$ est une chaîne de Markov à espace d'état E; notons par $\mathbf{p} = (p_{ij})_{i,j\in E}$ sa matrice de transition associée,

$$p_{ij} := \mathbb{P}(J_{m+1} = j \mid J_m = i),\ i, j \in E,\ m \in \mathbb{N}.$$

Pour tout $i, j \in E$, nous pouvons définir la loi conditionnelle de séjour

$$f_{ij}(k) := \mathbb{P}(X_{n+1} = k \mid J_n = i, J_{n+1} = j) = q_{ij}(k)/p_{ij} = q_{ij}(k)/\left(\sum_{l=1}^{\infty} q_{ij}(l)\right),\ k \in \mathbb{N}. \tag{3}$$

Il est évident que

$$q_{ij}(k) = p_{ij} f_{ij}(k). \tag{4}$$

Nous allons avoir besoin par la suite de considérer la fonction de répartition de la loi conditionnelle de séjour, notée par

$$F_{ij}(k) \;:=\; \mathbb{P}(T_{m+1} - T_m \leq k | J_m = i, J_{m+1} = j) = \sum_{t=1}^{k} f_{ij}(t). \tag{5}$$

Pour toute fonction de répartition $F(\cdot)$, la fonction de survie/fiabilité associée sera notée par $\overline{F}(k) := 1 - F(t)$.

Nous avons considéré un modèle semi-Markov général dont le noyau est donné par (4). Des cas particuliers peuvent être également considérés, en imposant aux lois du temps de séjour $f_{ij}(k)$ une dépendance seulement de l'état i, ou seulement de l'état j, ou ni de i ni de j. Par conséquent, nous pouvons définir les noyaux semi-markoviens :

Modèle particulier 1

$$\begin{aligned} q_{ij}(k) &:= p_{ij} f_{i\bullet}(k), \text{ où} \\ f_{i\bullet}(k) &= \mathbb{P}(T_{m+1} - T_m = k | J_m = i) = \sum_{v \in E} p_{iv} f_{iv}(k), \end{aligned} \tag{6}$$

Modèle particulier 2

$$\begin{aligned} q_{ij}(k) &:= p_{ij} f_{\bullet j}(k), \text{ où} \\ f_{\bullet j}(k) &= \mathbb{P}(T_{m+1} - T_m = k | J_{m+1} = j), \end{aligned} \tag{7}$$

Modèle particulier 3

$$\begin{aligned} q_{ij}(k) &:= p_{ij} f(k), \text{ où} \\ f(k) &= \mathbb{P}(T_{m+1} - T_m = k). \end{aligned} \tag{8}$$

2 Estimation paramétrique

Nous allons considérer les lois $f_{ij}(k) = f_{ij}(k; \theta_{ij})$ dépendantes des paramètres inconnus $\theta_{ij} \in \mathbb{R}^{m_{ij}}$, où la dimension de l'espace des paramètres m_{ij} est connue ; aucune hypothèse est faite sur $(p_{ij})_{ij}$. À partir des données, nous voulons estimer p_{ij} et θ_{ij}, $i, j \in E$. Ainsi on a un problème d'estimation paramétrique avec les paramètres $(\theta_{ij}, p_{ij}, i, j \in E) \in \mathbb{R}^{\sum_{i,j \in E} m_{ij} + s^2}$ (s est le cardinal de E) et les contraintes $\sum_{j \in E} p_{ij} = 1, i \in E$.

Nous décrivons ici seulement le cas général, avec plusieurs trajectoires, censure au début et à la fin des trajectoires. Il faut noter que dans certaines applications, des formes particulières intéressantes peuvent être obtenues si nous considérons des cas particuliers de noyau du type (6), (7) ou (8).

Nous supposons qu'on observe L trajectoires semi-markoviennes, chacune de longueur n_l, $l = 1, \ldots, L$, censurées au début et à la fin, i.e.,

$$j_0^l, k_0^l, j_1^l, k_1^l, j_2^l, k_2^l, \ldots, j_{t^l}^l, k_{t^l}^l, j_{t^l+1}^l, k_{t^l+1}^l$$

avec $\sum_{i=0}^{t^l+1} k_i^l = n_l$, où $j_0^l, \ldots, j_{t^l+1}^l$ sont les états visités, k_0^l le premier temps de séjour, censuré à droite, $k_{t^l+1}^l$ le dernier temps de séjour, censuré à droite, et $k_1^l, \ldots, k_{t^l}^l$ les autres temps complets (pas censurés).
Il faut remarquer que le premier temps de séjour, k_0^l, est bien censuré à droite et non à gauche comme on pourrait le penser, car nous observons la valeur k_0^l (valeur de censure) et nous savons que le vrai temps de séjour (non observé) a une valeur plus grande que celle que nous observons.

La vraisemblance est donnée par

$$\begin{aligned}
& L(p_{uv}, \theta_{uv}; u, v \in E) \\
= \; & \prod_{i \in E} \mu_i^{N_i^{start}(L)} \prod_{u,v \in E} p_{uv}^{N_{uv}(L, n_{1:L})} \prod_{u,v \in E} \prod_{k=1}^{\max_l(n_l)} f_{uv}(k; \theta_{uv})^{N_{uv}(k; L, n_{1:L})} \\
& \times \prod_{u,v \in E} \prod_{k=1}^{\max_l(n_l)} \overline{F}_{uv}(k; \theta_{uv})^{\overline{N}_{uv}^b(k;L)} \prod_{u \in E} \prod_{k=1}^{\max_l(n_l)} \overline{F}_{u\bullet}(k; \theta_{uv}, v \in E)^{\overline{N}_{u\bullet}^e(k;L)},
\end{aligned}$$

où nous avons introduit les processus de comptage suivants :

$$\begin{aligned}
N_{ij}(L, n_{1:L}) &= \sum_{l=1}^{L} \sum_{m=1}^{N^l(n_l)-1} \mathbb{1}_{\{J_m^l = i; J_{m+1}^l = j\}}, \\
N_{i\bullet}(L, n_{1:L}) &= \sum_{m=1}^{N^l(n_l)-1} \mathbb{1}_{\{J_m^l = i\}}, \\
N_{ij}(k; L, n_{1:L}) &= \sum_{m=1}^{N^l(n_l)-1} \mathbb{1}_{\{J_m^l = i; J_{m+1}^l = j; T_{m+1}^l - T_m^l = k\}}, \\
\overline{N}_{ij}^b(k; L) &= \sum_{l=1}^{L} \mathbb{1}_{\{J_0^l = i; J_1^l = j; T_1^l - T_0^l > k\}}, \\
\overline{N}_{i\bullet}^e(k; L) &= \sum_{l=1}^{L} \mathbb{1}_{\{J^l_{T^l_{N^l(n_l)}} = i, X^l_{T^l_{N^l(n_l)+1}} > k\}}, \\
N_i^{start}(L) &= \sum_{l=1}^{L} \mathbb{1}_{\{J_0^l = i\}},
\end{aligned}$$

où

$$N^l(n_l) = \max\{m \in \mathbb{N} \mid T^l_m \leq n_l\}$$

est le processus de comptage du nombre de sauts dans $[1; n_l]$ de la trajectoire l.

Les EMV sont obtenus en résolvant le problème

$$\begin{aligned}
&\underset{p_{uv},\theta_{uv};u,v\in E}{\operatorname{argmax}} \left(l(p_{uv},\theta_{uv};u,v\in E)\right) \\
= &\left(\underset{p_{uv},\theta_{uv};v\in E}{\operatorname{argmax}} \left(\sum_{v\in E} N_{uv}(L,n_{1:L})\log(p_{uv}) + \sum_{v\in E}\sum_{k=1}^{\max_l(n_l)} N_{uv}(k;L,n_{1:L})\log(f_{uv}(k;\theta_{uv}))\right.\right. \\
&+\sum_{v\in E}\sum_{k=1}^{\max_l(n_l)} \overline{N}^b_{uv}(k;L)\log(\overline{F}_{uv}(k;\theta_{uv})) \\
&\left.\left.+\sum_{k=1}^{\max_l(n_l)} \overline{N}^e_{u\bullet}(k;L)\log\left(1-\sum_{m=1}^{k}\sum_{v\in E} p_{uv}f_{uv}(m;\theta_{uv})\right)\right)_{u\in E}\right),
\end{aligned}$$

où nous avons utilisé le fait que la maximisation est faite séparément pour chaque $u \in E$.

3 Le package R associé

Un package R (cf. Barbu et. al [2]) a été développé pour mettre en place l'estimation et la simulation de modèles de Markov et semi-Markov. Pour le cas semi-markovien nous avons considéré le cas de l'estimation paramétrique ou non-paramétrique, sans censure, avec censure au début et/ou à la fin de la séquence, avec une ou plusieurs trajectoires. L'estimation non-paramétrique concerne les lois des temps de séjour, qui ont un support infini et pour lesquelles aucune hypothèse est faite. Pour l'estimation paramétrique, plusieurs lois discrètes sont considérées (uniforme, géométrique, poisson, weibull discret, binomiale négative). La figure suivante décrit ce package.

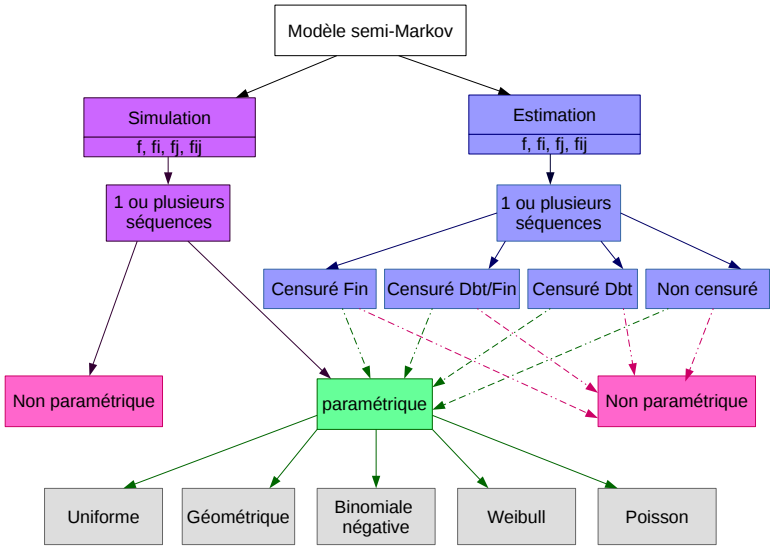

FIGURE 1 – La structure du package

Bibliographie

[1] Barbu, V. S., Bérard, C., Cellier, D., Sautreuil, M. et Vergne, N. (2017), Parametric estimation of semi-Markov chains, soumis.
[2] Barbu, V. S., Bérard, C., Cellier, D., Sautreuil, M. et Vergne, N. (2017), Semi-Markov models - an `R` package, en cours.
[3] Barbu, V. S. et Limnios, N. (2008), *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications - Their use in Reliability and DNA Analysis*, Lecture Notes in Statistics, vol. 191, Springer, New York.
[4] Barbu, V. S. et Limnios, N. (2006), Empirical estimation for discrete time semi-Markov processes with applications in reliability, *Journal of Nonparametric Statistics*, 18 (4), 483–498.
[5] Trevezas, S. et Limnios, N. (2011), Exact MLE and asymptotic properties for nonparametric semi-Markov models, *Journal of Nonparametric Statistics*, 23 (3), 719–739.