Tests d'égalité en loi pour des panels de processus semi-Markoviens : application en analyse sensorielle

Cindy Frascolla & Hervé Cardot
Institut de Mathématiques de Bourgogne, UMR CNRS 5584, Université de Bourgogne, 21000 Dijon, France

Résumé. Ce travail, motivé par une application en analyse sensorielle, s'intéresse à des tests d'hypothèses pour des panels de processus semi-Markoviens (PSM). En 2018, Lecuelle *et al.* ont proposé de modéliser les séquences des sensations perçues au cours de la dégustation d'un produit à l'aide de PSM. Un test statistique, basé sur le rapport de vraisemblance, a été construit et étudié par simulations (Frascolla *et al.* 2020) afin de déterminer si deux produits testés sont perçus différemment. Nous considérons des panels constitués de n trajectoires et deux modèles d'observation pour chaque trajectoire. Nous nous intéressons à des PSM composés d'un nombre aléatoire de transitions pour le premier modèle et à des PSM absorbants pour le second. Nous montrons la convergence asymptotique des estimateurs du maximum de vraisemblance des paramètres des PSM lorsque n tend vers l'infini et déterminons la distribution asymptotique du rapport de vraisemblance.

Mots-clés. Analyse sensorielle, Estimateur du maximum de vraisemblance, Processus semi-Markoviens, Statistique asymptotique, Test du rapport de vraisemblance

Abstract. This work, motivated by an application in sensory analysis, focuses on hypothesis testing for semi-Markov processes (SMP). In 2018, Lecuelle *et al.* have considered SMP to model the sequences of sensations perceived during the tasting of a product. A statistical test, based on the likelihood ratio, has been built and studied by simulations (Frascolla *et al.* 2020) in order to determine if two tested products are felt differently. We consider panels made up of n trajectories and two observation designs for each trajectory. We consider SMP composed of a random number of transitions for the first model and absorbing SMP for the second. We prove the asymptotic convergence of the estimators of the maximum likelihood of the parameters of the SMP when n tends to infinity and we determine the asymptotic distribution of the likelihood ratio.

Keywords. Sensory analysis, Maximum likelihood estimator, Semi-Markov processes, Asymptotic statistics, Likelihood ratio test

1 Introduction

L'analyse sensorielle a pour objectif de mieux comprendre les préférences des consommateurs. Lors d'une étude d'analyse sensorielle, différents produits d'une même catégorie

sont testés et les sujets indiquent la séquence des sensations perçues au cours du temps parmi une liste de descripteurs. En 2018, Lecuelle *et al.* ont proposé de modéliser ces données par des PSM pour prendre en compte la dynamique du modèle (les transitions d'un descripteur vers un autre) et les temps de séjour associés.

Les processus semi-Markoviens sont une généralisation des chaînes de Markov et permettent de considérer des modèles plus flexibles pour la loi des temps de séjour. Les livres de Limnios et Oprişan (2001) et de Barbu et Limnios (2008) présentent la théorie des processus semi-Markoviens et leur application en fiabilité et analyse de l'ADN.

Une des principales questions en analyse sensorielle est de déterminer si deux produits testés sont différents. Pour répondre à cette question, un test statistique basé sur le rapport de vraisemblance a été proposé dans Frascolla *et al.* (2020) avec trois approches pour déterminer la zone de rejet : deux approches basées sur des techniques de ré-échantillonnage (bootstrap paramétrique et permutations) et une approche basée sur la loi asymptotique du rapport de vraisemblance en supposant de manière intuitive qu'elle suivait une loi du χ^2, ce qui a été vérifié sur des simulations.

Nous montrons dans ce travail la consistance des estimateurs du maximum de vraisemblance (EMV) et leur normalité asymptotique et nous en déduisons la loi asymptotique du rapport de vraisemblance sous l'hypothèse d'égalité des lois.

2 Modèle et notations

2.1 Définition des processus semi-Markoviens

Soit $Z = (Z_t)_{t\in\mathbb{R}_+}$ un processus stochastique à valeur dans $E = \{1, \dots, D\}$ avec $D < +\infty$. Soient $J = (J_n)_{n\in\mathbb{N}}$ la suite des états visités par Z et $X = (X_n)_{n\in\mathbb{N}^*}$ la suite des temps de séjour associés. Nous supposons que le processus $(J_n, X_n)_{n\geq 1}$ vérifie la propriété de Markov, pour $t \in T = [0, +\infty[$, $\ell \in E$ et $j \neq \ell$, $\mathbb{P}(J_{n+1} = j, X_{n+1} \leq t \mid J_0, \dots, J_n, X_1, \dots, X_n) = \mathbb{P}(J_{n+1} = j, X_{n+1} \leq t \mid J_n)$.
Le processus donnant l'état visité à chaque instant t est appelé PSM (voir par exemple Limnios et Oprişan, 2001). La loi du PSM est caractérisée par sa distribution initiale $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)$ avec $\alpha_j = \mathbb{P}(J_0 = j)$, $j = 1, \dots, D$. La matrice de transition de la chaîne de Markov $(J_n)_{n\geq 1}$, notée $\mathbf{P}$, est constituée des éléments $p_{ij} = \mathbb{P}(J_n = j \mid J_{n-1} = i)$, pour $i \neq j \in E \times E$ et $p_{ii} = 0$ pour tout $i \in E$. On suppose que les distributions des temps de séjour appartiennent à une famille paramétrique de densité. Pour $i \neq j \in E \times E$, on note $f(t; \theta_{ij})$ la densité du temps de séjour de $X_n | J_{n-1} = i, J_n = j$ avec $\theta_{ij} \in \mathbb{R}^d$. On note $\boldsymbol{\theta} = (\theta_{ij}, i \neq j \in E \times E)$. Ainsi, la distribution du PSM Z est caractérisée par le vecteur des paramètres $(\boldsymbol{\alpha}_0, \mathbf{P}_0, \boldsymbol{\theta}_0)$.

2.2 Les différents protocoles d'observation considérés

On considère des panels constitués de n trajectoires indépendantes, $S_1, \ldots, S_n$, issues du même processus semi-Markovien de paramètre $(\boldsymbol{\alpha}_0, \mathbf{P}_0, \boldsymbol{\theta}_0)$. Une séquence S est constituée des différents états visités par le processus et des temps de séjour associés. Les données d'analyse sensorielle nous amènent à considérer deux protocoles d'observations. Pour le premier, chaque séquence est constituée d'un nombre aléatoire de transitions et on note M le nombre aléatoire de transitions. On suppose que M est strictement positif, d'espérance finie et vérifie les propriétés d'un temps d'arrêt. Pour le second, on suppose que l'espace d'état contient un état absorbant et chaque séquence s'arrête une fois qu'elle a atteint cet état absorbant. On suppose que l'état absorbant est accessible et on ré-ordonne les états de E de sorte que l'état absorbant soit le dernier état, c'est-à-dire l'état D. On suppose de plus que le processus débute dans un état non absorbant. La variance et l'espérance du nombre de transitions jusqu'à absorption pour des chaînes de Markov absorbantes sont finies (voir Kemeny et Snell (1976) par exemple).

3 Convergence asymptotique de l'EMV

La vraisemblance des n séquences s'écrit $\mathcal{L}(S_1, \ldots, S_n; \boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta}) = \prod_{l=1}^{n} \mathcal{L}(S_l; \boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta})$ où $\mathcal{L}(S_\ell; \boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta})$ est la vraisemblance de la séquence ℓ. La valeur moyenne de la log-vraisemblance sur les n trajectoires $S_1, \ldots, S_n$ est définie par :

$$\widehat{Q}(\boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{\ell=1}^{n} \ln\left(\alpha_{j_0^{(\ell)}}\right) + \widehat{Q}_{\mathbf{P}}(\mathbf{P}) + \widehat{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{1}$$

avec $\widehat{Q}_{\mathbf{P}}(\mathbf{P}) = \frac{1}{n} \sum_{\ell=1}^{n} \sum_{i \in E} \sum_{\substack{j=1 \\ j \neq i}}^{D} N_{ij}^{(\ell)} \ln(p_{ij})$ et $\widehat{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{\ell=1}^{n} \sum_{\substack{i,j \in E \\ j \neq i}} \sum_{k=1}^{N_{ij}^{(\ell)}} \ln\left(f(x_{ij}^{(\ell,k)}; \theta_{ij})\right)$ où $N_{ij}^{(\ell)}$ est le processus de comptage donnant le nombre de transitions de i vers j pour une séquence ℓ et $x_{ij}^{(\ell,k)}$ est le temps de séjour dans l'état i avant d'aller dans l'état j pendant la visite numéro k.

L'étude des propriétés asymptotiques de l'EMV peut être effectuée en trois parties en analysant séparément chaque ensemble de paramètres car la maximisation de la log-vraisemblance en $(\boldsymbol{\alpha}, \mathbf{P}, \boldsymbol{\theta})$ s'effectue de manière indépendante pour chaque ensemble de paramètres (1). On note $\widehat{\boldsymbol{\alpha}}$, $\widehat{\mathbf{P}}$ et $\widehat{\boldsymbol{\theta}}$ les EMV de $\boldsymbol{\alpha}$, $\mathbf{P}$ et $\boldsymbol{\theta}$. L'estimateur $\widehat{\boldsymbol{\alpha}}$ est simplement l'EMV pour une distribution multinomiale (voir Trevezas 2011 par exemple). La difficulté liée à l'estimation des paramètres $\mathbf{P}$ et $\boldsymbol{\theta}$ provient du fait que les sommations s'effectuent sur un nombre aléatoire d'indices, qui n'est pas nécessairement indépendant des variables aléatoires étudiées. Il faut également que les conditions classiques assurant la convergence

des estimateurs du maximum de vraisemblance soient vérifiées (conditions sur la loi des temps de séjour et conditions sur les p_{ij} et α_i).

Nous démontrons ensuite la consistance des EMV et leur normalité asymptotique en utilisant des outils statistiques (*"continuous mapping theorem"*, loi forte des grands nombres et identité de Wald par exemple) et les théorèmes de Newey et McFadden (1994).

4 Tests d'égalité en loi de deux processus

On souhaite maintenant comparer deux échantillons. Pour cela, on dispose de deux échantillons de taille n_1 et n_2 issus de deux processus semi-Markoviens Z^1 et Z^2 caractérisés par les vecteurs de paramètres $(\boldsymbol{\alpha}^1, \mathbf{P}^1, \boldsymbol{\theta}^1)$ et $(\boldsymbol{\alpha}^2, \mathbf{P}^2, \boldsymbol{\theta}^2)$. On souhaite tester l'hypothèse d'égalité des lois de ces deux processus. Pour cela, nous avons deux stratégies. La première repose sur un test du rapport de vraisemblance et la deuxième sur la statistique de Wald. Nous ne considérons pas les paramètres des probabilités initiales dans la suite car tester leur égalité par un test du rapport de vraisemblance relève d'outils classiques non présentés ici. Les hypothèses du test sont $H_0 : (\mathbf{P}^1, \boldsymbol{\theta}^1) = (\mathbf{P}^2, \boldsymbol{\theta}^2)$ et $H_1 : (\mathbf{P}^1, \boldsymbol{\theta}^1) \neq (\mathbf{P}^2, \boldsymbol{\theta}^2)$.
La statistique de test basée sur le rapport de vraisemblance, notée LR, est le quotient de la vraisemblance sous l'hypothèse nulle et de la vraisemblance sous l'hypothèse alternative. On peut noter que LR ne dépend pas du paramètre α grâce à la structure multiplicative de la vraisemblance. La statistique de test utilisant un test de Wald, notée $W^{p,\theta}_{n_1,n_2}$, est basée sur la distribution asymptotique des EMV sous l'hypothèse nulle.
Nous montrons, sous des hypothèses classiques pour les lois des temps de séjour pour le maximum de vraisemblance, que $-2\ln(LR)$ et $W^{p,\theta}_{n_1,n_2}$ convergent en loi vers une loi du χ^2 avec $D(D-2) + D(D-1)d$ degrés de liberté en l'absence d'état absorbant et $(D-1)(D-2) + (D-1)^2 d$ degrés de liberté quand il y a un état absorbant.

Bibliographie

Barbu, V. S. and Limnios, N. (2008). Semi-Markov chains and hidden semi-Markov models toward applications : their use in reliability and DNA analysis. *New York: Springer Science + Business Media.*
Frascolla, C., Lecuelle, G., Cardot, H., and Schlich, P. (2020). Two sample tests for semi-Markov processes with parametric sojourn time distributions: an application in sensory analysis. Article en révision
Kemeny, J.G, Snell, J.L. (1976). *Finite Markov Chains.* Springer-Verlag, New York-Heidelberg.
Lecuelle, G., Visalli, M., Cardot, H. and Schlich, P. (2018). Modeling temporal dominance of sensations with semi-Markov chains. *Food Quality and Preference* 67, 59–66.
Limnios, N. et G. Oprişan (2001). *Semi-Markov Processes and Reliability.* Statistics for Industry and Technology. Birkhäuser Boston, Inc., Boston, MA.
Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Vol. IV, pages 2111–2245. North-Holland, Amsterdam.
Trevezas, S. and Limnios, N. (2011). Exact MLE and asymptotic properties for nonparametric semi-Markov models. *Journal of Nonparametric Statistics* , 23, 952-958.