Comparaison de trajectoires qualitatives avec des chaînes semi-Markoviennes : une application en analyse sensorielle

Cindy Frascolla [1] & Guillaume Lecuelle [2] & Hervé Cardot [1]
& Michel Visalli [2] & Pascal Schlich [2]

[1] *Institut de Mathématiques de Bourgogne, UMR CNRS 5584, Université de Bourgogne, 21000 Dijon, France*
[2] *Centre des Sciences du Goût et de l'Alimentation, UMR AgroSup Dijon-CNRS-INRA-Université de Bourgogne, 21000 Dijon, France.*

Résumé.
L'analyse sensorielle est très utilisée dans l'industrie agroalimentaire pour développer des nouveaux produits. La Dominance Temporelle des Sensations (DTS) est une méthode d'analyse sensorielle, utilisant une liste de descripteurs, qui permet d'indiquer comment les sensations ressenties lors de la dégustation d'un produit changent au cours du temps. En 2018, Lecuelle et al. ont introduit des chaînes semi-Markoviennes aux données DTS. Pour comparer deux produits lors d'études DTS, on introduit dans ce travail un test statistique basé sur le test du rapport de vraisemblance entre deux modèles semi-Markoviens. Pour construire la zone de rejet, trois approches sont évaluées. Une première basée sur le bootstrap paramétrique, une seconde basée sur les tests de permutation et une troisième qui repose sur la loi asymptotique du rapport de vraisemblance. Ces approches sont comparées à partir de simulations et de tests réalisés sur des jeux de données réels de dégustations de chocolats et de fromages.

Mots-clés. Analyse sensorielle, Chaînes semi-Markoviennes, Loi gamma, Dominance Temporelle des Sensations (DTS), Statistique des processus à valeurs qualitatives, Test du rapport de vraisemblance

Abstract. Sensory analysis is very useful in food industry to develop new products. Temporal Dominance of Sensations (TDS) is a method of sensory analysis, using a list of attributes, which permits to indicate how sensations felt during the tasting of a product change over time. In 2018, Lecuelle and al introduced semi-Markov chains to TDS data. To compare two products in TDS studies, we introduce in this work a statistical test based on the likelihood ratio test between two semi-Markov models. To build the critical region, three approaches are evaluated. A first one based on the parametric bootstrap, a second one on permutations and a third one on the asymptotic law of the likelihood ratio. These approaches are compared on simulated data and real datasets which consist in tastings of chocolates and cheeses.

Keywords. Sensory analysis, Semi-Markov Chains (SMC), Gamma law, Temporal Dominance of Sensations (TDS), Statistics for continuous time qualitative processes, Likelihood ratio test

1 Introduction

L'analyse sensorielle, dont l'objectif est de mieux comprendre les préférences des consommateurs, est un outil indispensable à la mise au point de nouveaux produits. C'est un ensemble de méthodes visant à mesurer des perceptions sensorielles.

Plusieurs méthodes d'analyse sensorielle sont utilisées et étudiées dont la Dominance Temporelle des Sensations qui a été développée au Centre des Sciences du Goût et de l'Alimentation (CSGA) de Dijon par Pineau et al (2009). Son principe est le suivant : lors d'une dégustation, les juges ont une liste de descripteurs et ils choisissent celui qui leur semble dominant, c'est à dire celui qui « attire le plus leur attention » et ceci tout au long de la dégustation. Un descripteur est considéré dominant jusqu'à ce qu'un autre descripteur soit dominant. La période pendant laquelle un descripteur est considéré comme dominant est appelée durée de dominance. A chaque instant, un seul descripteur peut-être choisi.

Pour décrire les données DTS, plusieurs techniques sont utilisées comme les bandplots individuels ou les courbes DTS (Pineau et al. 2009). Les bandplots individuels représentent les descripteurs cités par chaque juge et les temps de dominance associés. Les courbes DTS représentent l'évolution du taux de dominance des descripteurs au cours du temps. Les bandplots individuels sont donc une représentation des comportements individuels alors que les courbes DTS sont une représentation du comportement du panel.

Aucune de ces techniques ne prend explicitement en compte la dépendance temporelle des données DTS. Pour remédier à cela, Franczak et al. (2015) proposent de les modéliser à l'aide de chaînes de Markov. Or, la durée des temps de séjour suit une loi géométrique dans le cas de chaînes de Markov. Cette loi n'est pas adaptée aux données DTS, voir Lecuelle et al. (2018), qui ont proposé de modéliser les trajectoires par des chaînes semi-Markoviennes (SMC). Les livres de Limnios et Oprişan (2001) et de Barbu et Limnios (2008) présentent la théorie des modèles semi-Markoviens et leur application en fiabilité et analyse de l'ADN. L'utilisation d'un modèle de mélange sur des chaînes semi-Markoviennes par Cardot et al. (2018) a permis de mettre en évidence l'existence de différents groupes de consommateurs lors de dégustations DTS.

Lors d'une étude d'analyse sensorielle, plusieurs juges testent différents produits d'une même catégorie. Aucun outil ne permet actuellement de comparer rigoureusement si les produits testés correspondent à des séquences de même loi ou non. Le but de ce travail est de construire un test statistique permettant de comparer plusieurs produits d'une étude DTS modélisés par des modèles semi-Markoviens. Le test statistique construit est basé sur le test du rapport de vraisemblance.

2 Modèle et notations

2.1 Définition des chaînes semi-Markoviennes

Soit $(J_p)_{p\geq 1}$ une chaîne de Markov homogène à valeurs dans un espace d'états fini, $E = \{1, \ldots, D\}$, de matrice de transition $\mathbf{P}$, constituée des éléments $P_{\ell j} = \mathbb{P}\left(J_{p+1} = j | J_p = \ell\right)$, $\ell, j \in E$. Notons $(X_p)_{p\geq 1}$ la suite des temps de séjour (aléatoires) dans les états visités par la chaîne $(J_p)_{p\geq 1}$. Pour $j \neq \ell$, on définit $\Phi_{\ell j}(t) = \mathbb{P}\left[X_p \leq t \mid J_p = \ell, J_{p+1} = j\right]$, la fonction de répartition du temps de séjour sachant l'état courant visité par la chaîne et son état futur. Nous supposons que le processus $(J_p, X_p)_{p\geq 1}$ vérifie la propriété de Markov, pour $t \in T = [0, +\infty[$, $\ell \in E$ et $j \neq \ell$,

$$\mathbb{P}\left[J_{p+1} = j, X_p \leq t \mid J_p = \ell, J_{p-1}, \cdots, J_1, X_{p-1}, \cdots, X_1\right] = P_{\ell j}\ \Phi_{\ell j}(t).$$

Le processus $(J_p, X_p)_{p\geq 1}$ est un processus de renouvellement de Markov tandis que le processus donnant l'étant visité à chaque instant t est appelé processus semi-Markovien (voir par exemple Limnios et Oprişan, 2001). Pour des raisons d'identifiabilité, nous supposons que $P_{jj} = 0$, pour tout $j \in E$. Finalement, pour caractériser complètement la loi de $(J_p, X_p)_{p\geq 1}$, nous définissons le vecteur $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$ des probabilités d'initialisation du processus, $\alpha_j = \mathbb{P}(J_1 = j),\ j \in E$.

Nous supposons, comme dans Cardot *et al.* (2018), que la loi des temps de séjour est une loi gamma de densité

$$f(t, a, \lambda) = \frac{t^{a-1}\lambda^a \exp(-\lambda t)}{\Gamma(a)}, \quad t \geq 0,$$

où $\Gamma(a)$ est la fonction gamma, a et λ sont des paramètres strictement positifs. De plus, nous considérons pour simplifier que la distribution des temps de séjour ne dépend que de l'état actuel, c'est à dire qu'il n'y a pas anticipation du prochain état de la chaîne. On a donc pour $t > 0$, $\ell \in E$, et $j \neq \ell$, $\Phi_{\ell j}(t) = \Phi_\ell(t)$.

La loi d'une chaîne de semi-Markov est donc complètement caractérisée par le paramètre multidimensionnel $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{P}, (a_\ell, \lambda_\ell)_{\ell\in E})$.

2.2 Modélisation des données DTS avec des chaînes semi-Markoviennes

Dans le cas de la modélisation des données DTS par des chaînes semi-Markoviennes, l'espace d'état E est composé de l'ensemble des descripteurs (voir Figure 1) du produit dégusté. L'état du système au cours de la n-ième sensation dominante est notée J_n et X_n est la durée correspondante. Sur l'exemple de modélisation des données DTS par des SMC (Figure 1), le goûteur a cité les descripteurs J_1="Crunchy", J_2="Cocoa" puis J_3="Melting" et enfin J_4="Sticky" avec des durées de sensation X_1, X_2, X_3 et X_4.

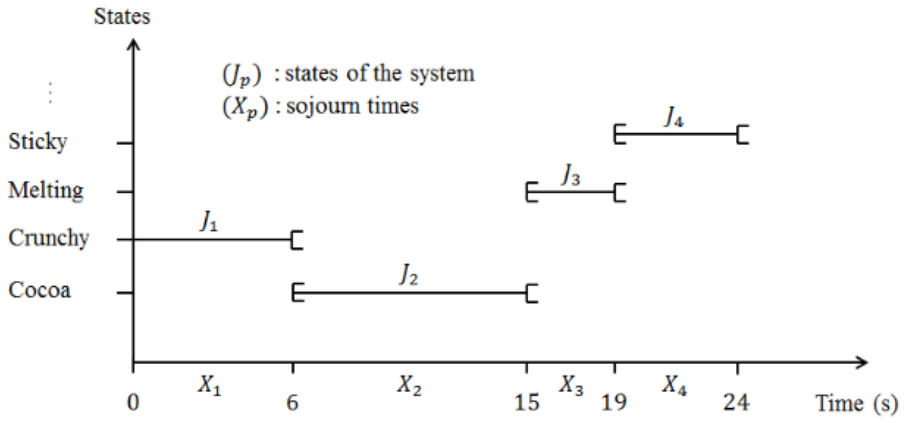

Figure 1: Modélisation d'une séquence DTS à l'aide d'une chaîne semi-Markovienne. (extrait de Cardot *et al.* 2018)

Chaque produit est dégusté par n "juges" (dégustateurs) et pour un juge i, on note S_i la séquence des sensations dominantes avec les temps de dominance,

$$S_i = (J_1^i, X_1^i, \ldots, J_{N(T_i)-1}^i, X_{N(T_i)-1}^i, J_{N(T_i)}^i, X_{N(T_i)}^i),$$

où $N(T_i)$ est le nombre de sensations ressenties par le juge i au cours de la dégustation. On suppose que $N(T_i) \geq 2$.

3 Vraisemblance et test du rapport de vraisemblance

En supposant que les séquences $S_1, S_2, ..., S_n$ sont indépendantes et issues d'une chaîne semi-Markovienne caractérisant le produit testé, la vraisemblance associée à $S_1, S_2, ..., S_n$ est $L_n(S_1, S_2, \ldots, S_n; \boldsymbol{\theta}) = \prod_{i=1}^n L(S_i; \boldsymbol{\theta})$, avec

$$L(S_i; \boldsymbol{\theta}) = \alpha_{J_1^i} \phi_{J_1^i}(X_1^i) \prod_{k=2}^{N(T_i)} \mathbf{P}_{J_{k-1}^i J_k^i} \phi_{J_k^i}(X_k^i)$$

où $\phi_\ell(x) = \int_0^x f(t, a_\ell, \lambda_\ell)dt$ est la fonction de répartition d'une loi gamma de paramètres $a = a_\ell$ et $\lambda = \lambda_\ell$.

Le problème de la comparaison de deux produits qui ont les mêmes descripteurs se modélise de la manière suivante : n_1 juges testent un premier produit, modélisé par une première chaîne semi-Markovienne de paramètres $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}^1, \mathbf{P}^1, (a_\ell^1, \lambda_\ell^1)_{\ell \in E})$, n_2 juges testent un autre produit modélisé par une deuxième chaîne semi-Markovienne de paramètres

$\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}^2, \mathbf{P}^2, (a_\ell^2, \lambda_\ell^2)_{\ell \in E})$. Le but du test étant de comparer ces deux produits, les hypothèses à confronter sont :

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$$
$$H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$$

Le test que nous considérons est basé sur le rapport de vraisemblance,

$$LR = \frac{\max_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^{n_1} L(S_i^1; \boldsymbol{\theta}) \times \prod_{j=1}^{n_2} L(S_j^2; \boldsymbol{\theta})}{\max_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta \times \Theta} \prod_{i=1}^{n_1} L(S_i^1; \boldsymbol{\theta}_1) \times \prod_{j=1}^{n_2} L(S_j^2; \boldsymbol{\theta}_2)}$$

où Θ est l'ensemble des valeurs possibles pour $\boldsymbol{\theta}$ et S_i^1 est la séquence de la dégustation du premier produit par le juge i et S_j^2 est la dégustation du deuxième produit par le juge j. Le rejet de l'hypothèse nulle s'effectue ensuite sur les petites valeurs de LR.

Pour construire la zone de rejet, trois approches sont comparées : une première approche basée sur la convergence asymptotique de la statistique de test vers une loi du χ^2 à p degrés de liberté, une seconde approche est basée sur le bootstrap paramétrique et une troisième sur une approche par permutations (voir Lehmann et Romano, 2005, pour une présentation de ces trois approches dans un cadre général).

La première approche repose sur les propriétés asymptotiques du test du rapport de vraisemblance. Si H_0 est vraie et si n_1 et n_2 tendent vers l'infini et sous des hypothèses de régularité, on a la convergence en loi $-2\ln(LR) \to \chi^2(p)$ où p est la différence entre le nombre de paramètres sous l'hypothèse nulle et sous l'hypothèse alternative. Dans notre cas, p est la dimension de Θ.

Le bootstrap paramétrique et les permutations consistent à approcher par Monte Carlo et simulations la loi du rapport de vraisemblance sous H_0. Pour cela, on regroupe les deux échantillons (on se place sous H_0) et on note $\widehat{\boldsymbol{\theta}}$ l'estimateur du paramètre $\boldsymbol{\theta}$ obtenu par maximum de vraisemblance. Dans le cas du bootstrap paramétrique, on simule ensuite un grand nombre B de fois ($B = 1000$), n_1 et n_2 trajectoires selon une chaîne semi-Markovienne de paramètre $\widehat{\boldsymbol{\theta}}$ et on calcule LR. Pour un risque de première espèce α, on rejette H_0 si la valeur de LR calculée sur l'échantillon observé est inférieure au quantile empirique d'ordre α de LR calculé sur les B expériences bootstrap. Pour l'approche par permutations (appelée aussi randomisation), on permute B fois les échantillons des deux produits en affectant pour chaque permutation, n_1 expériences au premier produit par tirage équiprobable sans remise dans les $n_1 + n_2$ expériences. Les n_2 autres expériences sont affectées au second produit. Le rejet ou non de H_0 s'effectue ensuite selon le même principe que le bootstrap paramétrique.

4 Illustration sur des données sensorielles

Des simulations, basées sur des jeux de données réels, ont été réalisées afin de comparer les trois approches présentées précédemment. Le premier jeux de données comporte $n_1 = 36$,

$n_2 = 36$ et $n_3 = 36$ expériences au cours desquelles trois chocolats ont été évalués selon $D = 10$ descripteurs. Le second jeu de données concerne des fromages, $n_1 = 665, n_2 = 665$ et $D = 10$ descripteurs. On compare le niveau théorique avec le niveau empirique sous H_0 et on évalue la puissance du test lorsque H_1 est vraie.

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*.

Cardot, H., Lecuelle, G., Visalli, M., and Schlich, P. (2018). Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data. `arXiv:1806.04420`.

Franczak, B. C., Browne, R. P., McNicholas, P. D., Castura, J. C. and Findlay, C. J. (2015). A Markov model for temporal dominance of sensations data. *In 11th Pangborn symposium*.

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.

Lehmann, E.L. and Romano, J. (2005). Testing statistical hypotheses. Third edition. Springer Texts in Statistics. Springer, New York.

Limnios, N. et G. Oprişan (2001). *Semi-Markov Processes and Reliability*. Birkhäuser, Boston.

Pineau, N., Schlich, P., Cordelle, S., Mathonniere, C., Issanchou, S., Imbert, A., Rogeaux, M., Etievant, P. and Koster, E. (2009). Temporal dominance of sensations: Construction of the tds curves and comparison with time-intensity. *Food Quality and Preference*, 20 (6), 450–455