Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data. arXiv:1806.04420v1 [stat.ME]

Hervé Cardot, Guillaume Lecuelle, Pascal Schlich, Michel Visalli

Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data

Hervé Cardot[a], Guillaume Lecuelle[b], Pascal Schlich[b], Michel Visalli[b]
a) Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
UBFC, F-21000 Dijon, France,
herve.cardot@u-bourgogne.fr
b) Centre des Sciences du Goût et de l'Alimentation,
AgroSup Dijon, CNRS, INRA, Université Bourgogne Franche-Comté,
F-21000 Dijon, France,
guillaume.lecuelle@inra.fr

June 13, 2018

Abstract

In food science, it is of great interest to get information about the temporal perception of aliments to create new products, to modify existing ones or more generally to understand the perception mechanisms. Temporal Dominance of Sensations (TDS) is a technique to measure temporal perception which consists in choosing sequentially attributes describing a food product over tasting. This work introduces new statistical models based on finite mixtures of semi-Markov chains in order to describe data collected with the TDS protocol, allowing different temporal perceptions for a same product within a population. The identifiability of the parameters of such mixture models is discussed. A penalty is added to the log likelihood to ensure numerical stability and consistency of the EM algorithm used to fit the parameters. The BIC criterion is employed for determining the number of mixture components. Then, the individual qualitative trajectories are clustered by considering the MAP criterion. A simulation study confirms the good behavior of the proposed estimation procedure. The methodology is illustrated on an example of consumers perception of a Gouda cheese and assesses the existence of several behaviors in terms of perception of this product.

Keywors : Bayesian information criterion; EM algorithm; gamma distribution; Identifiability; Markov renewal process; Model-based clustering; Penalized likelihood; Temporal dominance of sensations

1 Introduction

The development of food products is usually based on the measurement of product sensory perceptions from panels of consumers. Sensory perception while eating a food product has been acknowledged as a temporal process for 60 years (Neilson, 1957). Measuring

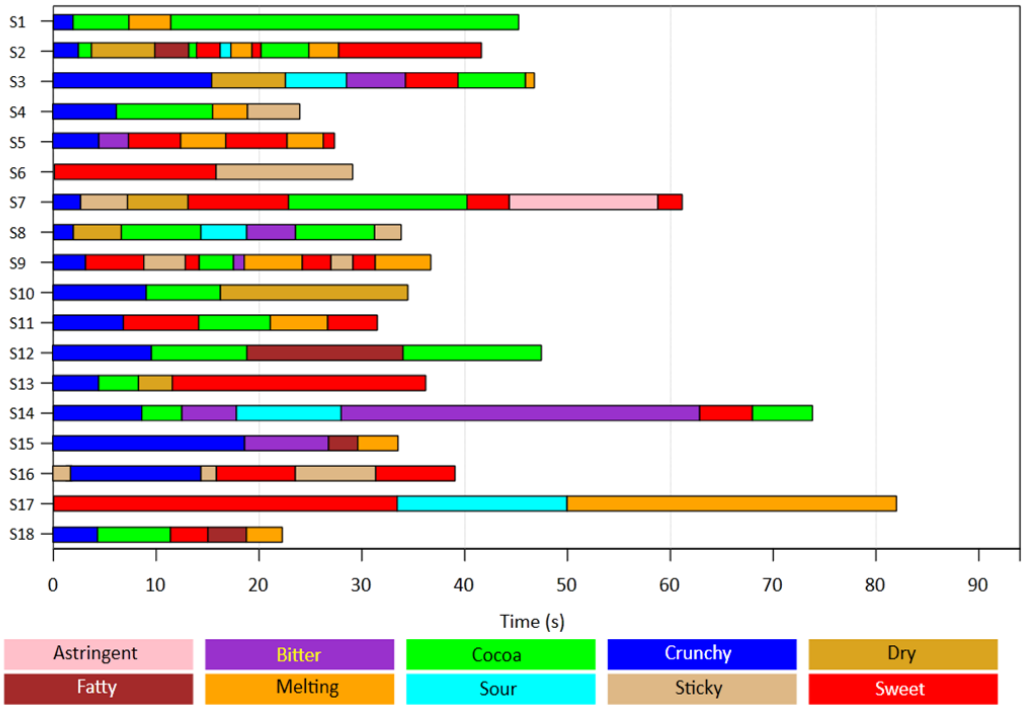

Figure 1: Tasting of a chocolate with 70% of cocoa by 18 panelists, denoted by S_1 to S_{18}, with 10 attributes. The bands represents the succession over time of the dominant attributes selected by each panelist while tasting this chocolate. The figure has been obtained by means of the TimeSens© software (www.timesens.com)

the temporal sensory perception is a complex task and different approaches have been developed in sensory science (see Hort et al. (2017)). Recently a technique called Temporal Dominance of Sensations (TDS) has been introduced by Pineau et al. (2009). A review on TDS can be found in Schlich (2017). The panelists have to describe the tasted product by choosing which attribute, among a list composed of about ten items, corresponds to the most striking perception at a given time. This task results in sequences of attributes with choices and time of the choices. When an attribute is selected as dominant, it is considered as dominant until the panelist select another dominant attribute. At each time only one attribute can be dominant. An example of such an experiment for a chocolate tasting is presented in Figure 1 with data represented as bandplots.

Some simple methods are currently used to describe such qualitative temporal data. Most of them rely on the observation of TDS curves, which consist in representing the evolution along time of the proportions of the dominant attributes at a panel level. Even if this statistical approach can be very informative, such a tool only provides a mean panel overview and no information about the individual variability. Some quantitative analysis are used as complement (Galmarini et al., 2017) but these methods only consider dominance durations (the time spent as dominant for each attribute). None of

these approaches takes into account the whole complexity of TDS data: choices of dominant attribute, order of the choices and dominance durations that are sojourn times in the successive dominant attributes. Recently, Franczak et al. (2015) proposed to model TDS data with Markov chains. The Markov hypothesis, meaning that the probability of the next choice of dominant attribute only depends on the current dominant attribute, seems to be reasonable from a sensory perspective. However, the Markov hypothesis imposes strong restrictions on the sojourn time distribution which should be geometrically distributed when considering a discrete time process, or exponentially distributed when considering a continuous time process (see *e.g.* Norris (1998) for a general presentation of Markov chains). In a recent paper by Lecuelle et al. (2018), it has been proposed to model TDS data with semi-Markov chains (SMC) in order to better fit to the data by allowing arbitrarily distributed sojourn times. SMC, or Markov renewal processes, which have been introduced more than sixty years ago (Lévy, 1956; Smith, 1955), are now widely used in numerous fields of science such as queuing theory, reliability and maintenance, survival analysis, performance evaluation, biology, DNA analysis, risk processes, insurance and finance or earthquake modeling (see *e.g.* Barbu and Limnios (2008) and references therein).

It has often been suggested by sensory scientists (Jaeger et al., 2017) that consumers form non homogeneous populations, so that modeling the panelists perception with the help of a mixture model can be of real interest. A mixture model (McLachlan and Peel, 2000; Melnykov and Maitra, 2010) is a probabilistic model enabling to represent the presence of sub populations within an overall population. Finite mixture models are widely used in numerous fields such as biology or economy because they offer to unsupervised classification a flexible method based on a rigorous model and understandable results. A general framework for model-based clustering is presented in Banfield and Raftery (1993) and in McNicholas (2016). Mixture models are commonly used with the Gaussian distribution because of its ability to fit to a lot of problems, but it can also be used with any parametric model. Mixtures of Markov chains have been used in different fields such as finance (Frydman, 2005), computer science (Song et al., 2009) or road traffic estimation (Lawlor and Rabbat, 2017). In continuous time and continuous response, Delattre et al. (2016) introduce mixtures of stochastic differential equations and use a classification rule based on estimated posterior probabilities to cluster growth curves. However, as far as we know, the present work is the first one that considers mixtures of semi-Markov processes.

The purpose of this article is to estimate mixtures of SMC, in discrete or continuous time, in order to model TDS data and to perform a segmentation of a sample of panelists into groups with similar perception. Identifiability is a crucial issue for mixture models (see Titterington et al. (1985)) and we show under general conditions that, when parametric models are considered for the distribution of sojourn times, the parameters of the model are identifiable up to label swapping. The estimation of the parameters is performed with the EM algorithm (McLachlan and Krishnan, 2008) in which a penalty may be added to ensure convergence. As explained in Chen et al. (2016) and Chen (2017), the likelihood is generally unbounded in case of mixtures of normal or gamma distributions and considering a penalized likelihood criterion leads to consistent and more stable

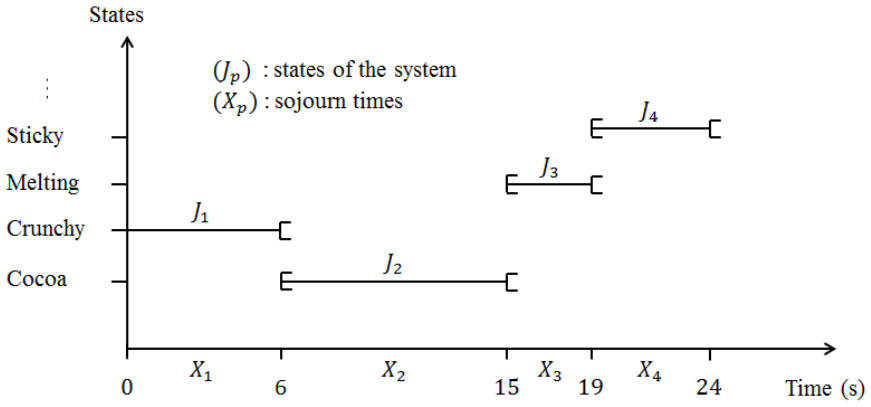

Figure 2: Modeling of sequence $S4$ (see Figure 1) with a Markov renewal process $(J_p, X_p)_{p\geq 1}$. The successive states chosen by the panelist are J_1 =Crunchy, J_2 =Cocoa, J_3 =Melting and J_4 =Sticky

estimates. The number of mixture components being generally unknown, an information criterion is employed to select the number of sub populations that should be considered. Then, the observed trajectories can be clustered using the maximum a posteriori (MAP) classification criterion (see Frühwirth-Schnatter (2006)).

The proposed method is illustrated on a dataset from the European Sensory Network (Thomas et al., 2017). This dataset includes TDS data for 4 Gouda cheeses tasted by 665 consumers according to 10 attributes. A mixture of SMC with gamma sojourn time distributions is adjusted to fit the data.

The article is organized as follows. Section 2 presents the mixture models and discusses the identifiability issue. Section 3 presents the EM algorithm employed for the estimation of the parameters of the mixture, the proportions and the number of components. Section 4 evaluates the performances of the statistical methods through a simulation study and Section 5 provides an illustration of the proposed method on cheese tasting data. Concluding remarks and discussion are given in Section 6.

2 Stochastic model and notations

2.1 Markov renewal processes and finite mixtures of Markov renewal processes

Consider a finite state homogeneous Markov chain $(J_p)_{p\geq 1}$, taking values in the finite state space $\mathcal{S} = \{1, \ldots, D\}$, with transition matrix $\mathbf{P}$, whose generic elements are $P_{\ell j} = \Pr[J_{p+1} = j | J_p = \ell]$, $\ell, j \in \mathcal{S}$. Consider also the random sequence $(X_p)_{p\geq 1}$ made by the successive sojourn times in the visited states. For each $p \geq 1$, X_p represents the sojourn time at state J_p and takes values in $T = 1, 2, \ldots$ if time, denoted by t, is discrete and in $T = [0, +\infty[$ if time is continuous. We denote by $\Phi_{\ell j}(t) = \Pr[X_p \leq t \mid J_p = \ell, J_{p+1} = j]$,

the cumulative distribution function of the sojourn time given the current and the next states of the random process $(J_p)_{p\geq 1}$. We suppose that the random process $(J_p, X_p)_{p\geq 1}$ satisfies the Markov property, for all $t \in T$ and $\ell, j \in \mathcal{S}$,

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

The process $(J_p, X_p)_{p\geq 1}$ is called a Markov renewal process, whereas the stochastic process giving the state of the system at every time $t \in T$ is called a semi-Markov process (see *e.g.* Barbu and Limnios (2008)). For identifiability reasons, it is also supposed that $P_{jj} = 0$, for all $j \in \mathcal{S}$, so that at each jump, the system cannot remain in the same state. Finally, to completely characterize the law of $(J_n, X_n)_{n\geq 1}$ we define the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$ of initialization probabilities

$$\alpha_j = \Pr\left[J_1 = j\right], \quad j \in \mathcal{S}. \tag{2}$$

The example given in Figure 2 describes the modeling of the 4^{th} TDS sequence of the dataset presented in Figure 1.

The distribution of the semi-Markov process $(J_n, X_n)_{n\geq 1}$ is completely characterized by the set of parameters $(\boldsymbol{\alpha}, \mathbf{P}, \Phi_{\ell j}, \ell, j \in \mathcal{S})$ and in the following its probability law is denoted by $\text{Law}\,(\boldsymbol{\alpha}, \mathbf{P}, \Phi_{\ell j}, \ell, j \in \mathcal{S})$.

Let us consider now G independent semi-Markov processes taking values in the same state space $\mathcal{S}$, and for $g = 1, \ldots, G$, the initialization vector of probabilities $\boldsymbol{\alpha}^g$, the transition matrix $\mathbf{P}^g$, and the cumulative distribution functions for the sojourn times $\Phi^g_{\ell j}(t),\ t \in T$. Denoting by $\pi_g > 0$, the probability of observing a Markov renewal process with parameters $\left(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi^g_{\ell j}, \ell, j \in \mathcal{S}\right)$, we consider the finite mixture process $(J^\pi_n, X^\pi_n)_{n\geq 1}$ whose law is given by

$$\sum_{g=1}^{G} \pi_g \text{Law}\left(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi^g_{\ell j}, \ell, j \in \mathcal{S}\right). \tag{3}$$

The following proposition states that a finite mixture of Markov renewal processes is a Markov renewal process.

Proposition 2.1 *The process $(J^\pi_p, X^\pi_p)_{p\geq 1}$ is a Markov renewal process with parameters*

$$\left(\sum_{g=1}^{G} \pi_g \boldsymbol{\alpha}^g, \sum_{g=1}^{G} \pi_g \mathbf{P}^g, \sum_{g=1}^{G} \pi_g \Phi^g_{\ell j}, \ell, j \in \mathcal{S}\right).$$

Proof *The proof is immediate. Consider the unobserved latent class variable Z, taking values in $\{1, \ldots, G\}$ and satisfying $\Pr[Z = g] = \pi_g$ for $g = 1, \ldots, G$. The law of $(J^\pi_p, X^\pi_p)_{n\geq 1}$ given $Z = g$ can be expressed as Law $\left(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi^g_{\ell j}, \ell, j \in \mathcal{S}\right)$. Thus,*

$$\alpha^\pi_j = \Pr\left[J^\pi_1 = j\right] = \sum_{g=1}^{G} \Pr\left[J^\pi_1 = j | Z = g\right] \Pr[Z = g] = \sum_{g=1}^{G} \pi_g \alpha^g_j.$$

We also clearly have that $(J_n^\pi)_{n\geq 1}$ *is a Markov chain, with transition probabilities,*

$$\mathbf{P}_{\ell j}^\pi = \Pr\left[J_{n+1}^\pi = j | J_n^\pi = \ell\right] = \sum_{g=1}^{G} \Pr\left[J_{n+1}^\pi = j | J_n^\pi = \ell, Z = g\right] \Pr[Z = g] = \sum_{g=1}^{G} \pi_g \mathbf{P}_{\ell j}^g,$$

and, for $t \in T$,

$$\Phi_{\ell j}^\pi(t) = \sum_{g=1}^{G} \Pr\left[X_n \leq t \mid J_{n+1} = j, J_n = \ell, Z = g\right] \Pr[Z = g] = \sum_{g=1}^{G} \pi_g \Phi_{\ell j}^g(t).$$

□

2.2 The identifiability issue

Identifiability of mixture models can be a complicated issue (see *e.g.* Teicher (1963), Yakowitz and Spragins (1968), Titterington et al. (1985) or Allman et al. (2009)). However, identifiability of the parameters of a stochastic model is a very important condition to ensure the convergence of estimation algorithms to a unique value. We consider here a parametric framework and we are interested in models defined by a family of distributions $\mathcal{F}(\boldsymbol{\Theta}) = \{\text{Law}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ where $\boldsymbol{\Theta} \subset \mathbb{R}^q$ is the parameter space and $\boldsymbol{\theta}$ is a vector of parameters characterizing the probability distribution. We consider convex combinations of probability laws in $\mathcal{F}(\boldsymbol{\Theta})$, $\sum_{g=1}^{G} c_g \text{Law}(\boldsymbol{\theta}_g)$, with $\sum_g c_g = 1$, $c_g > 0$ and $\boldsymbol{\theta}_g \in \boldsymbol{\Theta}$, for $g = 1, \ldots, G$.

Adopting the same definition as in Yakowitz and Spragins (1968), we say that the finite mixtures are identifiable in the family $\mathcal{F}(\boldsymbol{\Theta})$ if and only if the convex hull of $\mathcal{F}(\boldsymbol{\Theta})$ has the uniqueness representation property:

$$\sum_{g=1}^{G} c_g \text{Law}(\boldsymbol{\theta}_g) = \sum_{h=1}^{H} c'_h \text{Law}(\boldsymbol{\theta}_h) \tag{4}$$

implies $G = H$ and for each $g \in 1, \ldots, G$ there is some $h \in \{1, \ldots, G\}$ such that $c_g = c'_h$ and $\boldsymbol{\theta}_g = \boldsymbol{\theta}_h$.

Moreover, it has been proven in Yakowitz and Spragins (1968) that finite mixtures of a family $\mathcal{F}(\boldsymbol{\Theta})$ are identifiable if and only if the cumulative distribution functions of the elements are linearly independent.

We suppose from now that the family of distributions of the sojourn times is parametric, $\Phi_{\ell j}(t) = \Phi(t, \boldsymbol{\Gamma}_{\ell j})$, with $\boldsymbol{\Gamma}_{\ell j} \in \mathbb{R}^d$. Classical parametric distributions of sojourn times are the negative binomial distribution if time is discrete ($d = 2$), and exponential ($d = 1$) or gamma distributions ($d = 2$) if time is continuous. In our renewal Markov processes framework, a parameter $\boldsymbol{\theta}_g$ will be of the form $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}, \boldsymbol{P}, \boldsymbol{\Gamma}_{\ell j}, \ell, j \in S)$. It is shown in Teicher (1963) that finite mixtures of Gamma distributions are identifiable whereas it is proven in Yakowitz and Spragins (1968) that finite mixtures of exponential distributions as well as negative binomial distributions are also identifiable. More recently, it has been shown in Gupta et al. (2016), under technical assumptions, that

almost all finite mixtures of Markov chains with D states are identifiable provided that at least three transitions can be observed and the number of mixture components is not too large compared to the number of states, more precisely $D \geq 2G$.

As shown in the next proposition, the identifiability of mixtures of renewal Markov processes can be assessed under much more weaker conditions than mixtures of Markov chains because of their bi-dimensional nature. We can note that no particular condition on the number of mixture components or on the number of states is required in the following proposition.

Proposition 2.2 *Suppose that the family of sojourn time distributions is identifiable. Then all the finite mixtures from the family $\mathcal{F}(\mathbf{\Theta})$ can be identified when the law of the sequence $(J_1^\pi, X_1^\pi, J_2^\pi, X_2^\pi)$ drawn form a mixture of renewal Markov processes is known.*

In other words, this proposition tells us we will be able to identify the parameters of a finite mixture from $\mathcal{F}(\mathbf{\Theta})$ provided that we can observe at least one transition and the two first sojourn times.

Proof *We proceed in an iterative way to prove the proposition by considering the law of the different events to successively identify the mixture probabilities and the sojourn time parameters, the transition probabilities and finally the initialization probabilities.*

First consider the distribution of the sojourn times between the first and the second states, $\Phi_{\ell j}^\pi(t) = \Phi(t, \mathbf{\Gamma}_{\ell j}^\pi)$, for $j \neq \ell \in \mathcal{S}$ and $t \in T$,

$$\begin{aligned}\Phi(t, \mathbf{\Gamma}_{\ell j}^\pi) = \Pr[X_1^\pi \leq t | J_1 = \ell, J_2 = j] &= \sum_{g=1}^{G} \pi_g \Pr[X_1^\pi \leq t | J_1 = \ell, J_2 = j, Z = g] \\ &= \sum_{g=1}^{G} \pi_g \Phi(t, \mathbf{\Gamma}_{\ell j}^g). \end{aligned} \tag{5}$$

By hypothesis, the family of sojourn times distributions is identifiable and $\Pr[J_1 = \ell] > 0$, for $\ell \in \mathcal{S}$, each component $\Phi(t, \mathbf{\Gamma}_{\ell j}^\pi)$ of the family of sojourn time distributions is identifiable, so the whole family $\Phi(t, \mathbf{\Gamma}_{\ell j}^\pi), \ell, j \in \mathcal{S}$ is also identifiable. We can suppose now that $\pi_1, \ldots, \pi_G$ and $(\mathbf{\Gamma}_{\ell j}^g, \ell, j \in \mathcal{S})$, for $g = 1, \ldots, G$ are known.

Now consider the law of the events, for $\ell \in \mathcal{S}$,

$$\begin{aligned}\Pr[X_2^\pi \leq t | J_2^\pi = \ell] &= \sum_{g=1}^{G} \pi_g \Pr[X_2 \leq t | J_2 = \ell, Z = g] \\ &= \sum_{g=1}^{G} \pi_g \sum_{j=1}^{D} \Pr[X_2 \leq t, J_3 = j | J_2 = \ell, Z = g] \\ &= \sum_{g=1}^{G} \pi_g \sum_{j=1}^{D} \mathbf{P}_{\ell j}^g \Phi(t, \mathbf{\Gamma}_{\ell j}^g). \end{aligned} \tag{6}$$

From the linear independence of the functions $\Phi(t, \mathbf{\Gamma}^g_{\ell j})$, for $g = 1, \ldots, G$, $\ell = 1, \ldots, D$ and $j \neq \ell$, and from the fact that $\pi_1, \ldots, \pi_G$ and $(\mathbf{\Gamma}^g_{\ell j}, \ell, j \in \mathcal{S})$, for $g = 1, \ldots, G$ are known, we can deduced that the coefficients $\mathbf{P}^g_{\ell j}$ can be uniquely determined. We can now suppose that the transition matrices $\mathbf{P}_g$, $g = 1, \ldots, G$, have been identified and are known.

To finish, the proof, consider now the law of the events, for $\ell, j \neq \ell \in \mathcal{S} \times \mathcal{S}$, and $t \in T$,

$$\begin{aligned} \Pr\left[X^\pi_1 \leq t, J_1 = \ell, J_2 = j\right] &= \sum_{g=1}^{G} \pi_g \Pr[X_1 \leq t, J_1 = \ell, J_2 = j | Z = g] \\ &= \sum_{g=1}^{G} \pi_g \left(\alpha^g_\ell \mathbf{P}^g_{\ell j} \Phi(t, \mathbf{\Gamma}^g_{\ell j})\right) \end{aligned} \tag{7}$$

As before, all the components at the right-hand side of (7) are known, except the initialization probabilities α^g_ℓ, $g = 1, \ldots, G$ and $\ell \in \mathcal{S}$. Since the functions $\Phi(t, \mathbf{\Gamma}^g_{\ell j})$ are linearly independent, the parameters $\alpha^g_\ell \mathbf{P}^g_{\ell j}$ can be identified. The parameters $\mathbf{P}^g_{\ell j}$ being known, the initialization probabilities α^g_ℓ, for $g = 1, \ldots, G$ and $\ell \in \mathcal{S}$, are identifiable. □

3 Maximum likelihood estimation and model selection

Consider a sample of n independent sequences S_i, $i = 1, \ldots, n$, of a Markov renewal process. Each sequence S_i is observed for $t \leq T_i$. The number of visited states for sequence S_i is denoted by $N(T_i)$ and we suppose that $N(T_i) \geq 2$. Thus we have,

$$S_i = (J^i_1, X^i_1, \ldots, J^i_{N(T_i)-1}, X^i_{N(T_i)-1}, J^i_{N(T_i)}, X^i_{N(T_i)}), \quad i = 1, \ldots, n. \tag{8}$$

We suppose that $S_1, \ldots, S_n$ are drawn from a mixture of G semi-Markov processes whose law is given in (3) and we aim at estimating the parameters which characterize the law of the mixture: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, and $(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi^g_{\ell j}, \ell, j \in \mathcal{S})$, for $g = 1, \ldots, G$. We suppose in this Section that the number G of components is known.

3.1 The likelihood

Let us denote by $\phi_{\ell,j}(t)$ the density function (or the probability for discrete time) of the sojourn time at the current state, given the current state is ℓ and the future state is j. Denote by $L_g(S_i; \boldsymbol{\theta}_g)$ the likelihood of a sequence S_i drawn from the Markov renewal process with parameters $\boldsymbol{\theta}_g = \left(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi^g_{\ell j}, \ell, j \in \mathcal{S}\right)$. We get, by successive conditioning on the past values:

$$L_g(S_i; \boldsymbol{\theta}_g) = \alpha^g_{J^i_1} \prod_{k=2}^{N(T_i)} \mathbf{P}^g_{J^i_{k-1} J^i_k} \phi^g_{J^i_{k-1} J^i_k}(X^i_{k-1}) \left(\sum_{\ell=1}^{D} \mathbf{P}^g_{J^i_{N(T_i)} \ell} \phi^g_{J^i_{N(T_i)} \ell}\left(X^i_{N(T_i)}\right)\right). \tag{9}$$

If we do not suppose anymore that the mixture component from which unit i arises is known, the log likelihood under the mixture model of the trajectories $S_1, \ldots, S_n$ is

$$\ln L(S_1, \ldots, S_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left(\sum_{g=1}^{G} \pi_g L_g(S_i; \boldsymbol{\theta}_g) \right), \tag{10}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the set of parameters of the mixture model. A direct maximization of the log-likelihood (10), according to $\boldsymbol{\theta}$ is cumbersome and classical optimization algorithm are generally not suitable to deal with that kind of problem (see *e.g* McLachlan and Krishnan (2008)). The EM algorithm, presented below, is preferred because it allows the optimization procedure to be decomposed into two simple steps.

3.2 The EM algorithm

The Expectation Maximization (EM) algorithm is a very useful algorithm that has first been designed to perform maximum likelihood estimation for incomplete data problems (see Dempster et al. (1977) and McLachlan and Krishnan (2008)). It is an iterative optimization technique of the likelihood that can be very effective for estimating mixture models by considering the unknown mixture components as missing observations (see McLachlan and Peel (2000)).

Let us introduce the missing mixture component indicators, Z_i, for $i = 1, \ldots, n$, which are vectors with G elements, composed of 1 one and $G-1$ zeros and that indicates from which component of the mixture the trajectory S_i arises. In other words, if S_i has been generated by the g^{th} mixture component then $Z_{ig} = 1$ and $Z_{i\ell} = 0$ for $\ell \neq g$. The complete data log-likelihood can be written as follows:

$$\begin{aligned} \ln L_c(S_1, Z_1, \ldots, S_n, Z_n; \boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \ln \left(\pi_g L_g(S_i; \boldsymbol{\theta}_g) \right) \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \ln \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \ln L_g(S_i; \boldsymbol{\theta}_g). \end{aligned} \tag{11}$$

This function is easier to maximize, according to $\boldsymbol{\theta}$, than the log-likelihood function given in (10).

An initial value $\boldsymbol{\theta}^{(0)}$ of the parameters must be carefully chosen before starting the algorithm. The choice of the starting point can be of great importance and is discussed in Section 3.4. The EM algorithm proceeds iteratively according to the following scheme. Suppose an estimate of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^{(m-1)}$, has been calculated at step $m-1$, with $m \geq 1$.

E-step

The expectation (E) step consists in computing the expected log-likelihood of the complete data given the observed trajectories and the value of the parameters estimated

during the previous iteration. We define

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)}) = \mathrm{E}\left[\ln L_c\left(S_1, Z_1, \ldots, S_n, Z_n; \boldsymbol{\theta}\right) | S_1, \ldots, S_n, \boldsymbol{\theta}^{(m-1)}\right]$$
$$= \sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{Z}_{ig}^{(m)} \ln L_g\left(S_i; \boldsymbol{\theta}_g^{(m-1)}\right) + \sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{Z}_{ig}^{(m)} \ln \pi_g^{(m-1)}, \qquad (12)$$

with $\widehat{Z}_{ig}^{(m)} = \mathrm{E}[Z_{ig}|S_1, \ldots, S_n, \boldsymbol{\theta}^{(m-1)}]$, the conditional probability for S_i to be generated by the component g of a mixture model with parameters $\boldsymbol{\theta}^{(m-1)}$, where $\boldsymbol{\theta}^{(m-1)}$ is the value of the set of parameters computed at previous iteration. We get with Bayes theorem,

$$\widehat{Z}_{ig}^{(m)} = \Pr(Z_{ig} = 1 | S_i; \boldsymbol{\theta}^{(m-1)})$$
$$= \frac{\pi_g^{(m-1)} L_g(S_i; \boldsymbol{\theta}^{(m-1)})}{\sum_{j=1}^{G} \pi_j^{(m-1)} L_j(S_i; \boldsymbol{\theta}^{(m-1)})}. \qquad (13)$$

M-step

The maximization (M) step consists in updating the estimation of parameter $\boldsymbol{\theta}$ given the expected values of $\hat{Z}_{ig}$, for $g = 1, \ldots, G$ and $i = 1, \ldots, n$, by looking for the maximum, according to $\boldsymbol{\theta}$, of the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)})$ defined in (12). The estimated mixture probabilities π_g at step m are obtained by solving

$$\frac{\partial}{\partial \pi_g}\left[\sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \ln(\pi_g) + \lambda\left(\sum_{g=1}^{G} \pi_g - 1\right)\right] = 0, \qquad (14)$$

where λ is the Lagrange multiplier associated to the constraint $\sum_{g=1}^{G} \pi_g = 1$. We get the standard solution, $\pi_g^{(m)} = n^{-1} n_g^{(m)}$, with $n_g^{(m)} = \sum_{i=1}^{n} \widehat{Z}_{ig}^{(m)}$.

Then, the Markov chain transition matrix and the initialization probabilities $(\boldsymbol{\alpha}_g, \mathbf{P}_g, g = 1, \ldots, G)$ and the parameters related to the sojourn time distributions $(\Phi_{\ell j}^{g}, \ell, j \in \mathcal{S}, g = 1, \ldots, G)$ are updated by maximizing the first term at the right-hand side of equality (12). As explained in Chen (2017), it may be preferable to add a penalty term to the part of the likelihood related to the sojourn time parameters in order to ensure consistency of the procedure.

Once the algorithm has converged, model-based clustering of the observed sequences is performed by considering the maximum *a posteriori* (MAP) criterion, defined as follows: $\mathrm{MAP}(\hat{Z}_{ig}) = 1$ if $g = \arg\max_h(\hat{Z}_{ih})$ and $\mathrm{MAP}(\hat{Z}_{ig}) = 0$ otherwise.

3.3 The particular case of gamma distributed sojourn times with replications and no anticipation

In our sensory examples we have considered sojourn times distributed according to gamma distributions because they are flexible distributions with simple moments that

are able to fit many different shapes of sojourn time behaviors. Their densities depend on two parameters $a > 0$ and $\lambda > 0$ and are defined as follows,

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

with corresponding expected value a/λ and variance a/λ^2.

We suppose, as in Lecuelle et al. (2018), that the sojourn time distribution only depends on the current state,

$$\Pr[X_1^\pi \leq t | J_1 = \ell, J_2 = j, Z = g] = \Pr[X_1^g \leq t | J_1 = \ell] \tag{15}$$

so that there is no anticipation, in some sense, of the next dominant attribute. This assumption, which seems relevant in a food tasting context, also allows us to deal with moderate size samples by reducing significantly the number of parameters to be estimated. Indeed, with this simplification, we only require to estimate GDd parameters to characterize the sojourn time distributions instead of $GD(D-1)d$ in the more general setting studied in previous Section. Note that $d = 2$ in the particular case of gamma distributed sojourn times.

We consider n experiments with B independent replications. For each panelist i, we get B sequences S_i^b, $b = 1, \ldots, B$, observed for $t \leq T_i^b$ and denoted by,

$$S_i^b = (J_1^{i,b}, X_1^{i,b}, \ldots, J_{N(T_i^b)-1}^{i,b}, X_{N(T_i^b)-1}^{i,b}, J_{N(T_i^b)}^{i,b}, X_{N(T_i^b)}^{i,b}), \quad i = 1, \ldots, n. \tag{16}$$

The likelihood related to a statistical unit i with B independent replications drawn from a Markov renewal process with parameters $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}^g, \mathbf{P}^g, (a_{\ell g}, \lambda_{\ell g}), \ell \in \mathcal{S})$ is:

$$L_g(S_i^1, \ldots, S_i^B; \boldsymbol{\theta}_g) = \prod_{b=1}^{B} \alpha^g_{J_1^{i,b}} \phi^g_{J_1^{i,b}}(X_1^{i,b}) \prod_{k=2}^{N(T_i^b)} \mathbf{P}^g_{J_{k-1}^{i,b} J_k^{i,b}} \phi^g_{J_k^{i,b}}(X_k^{i,b}) \tag{17}$$

where $\phi_\ell^g(t)$ is the cumulative distribution function for a gamma random variable with parameters $a = a_{\ell g}$ and $\lambda = \lambda_{\ell g}$.

Thanks to the multiplicative structure of the likelihood (17) given the mixture component, the conditional expectation of the complete log-likelihood can be written as the sum of two distinct functions, where the first one only depends on the semi-Markov chains parameters $(\boldsymbol{\alpha}_g, \mathbf{P}_g, g = 1, \ldots, G)$ whereas the second one only depends on the sojourn time distributions $(a_{\ell g}, \lambda_{\ell g}, \ell \in \mathcal{S}, g = 1, \ldots, G)$. Thus, these two sets of parameters can be estimated separately by maximizing each part of the log-likelihood at each M-step. Introducing again Lagrange multipliers, this yields the standard solution for the transition probabilities estimators as well as the initialization probabilities:

$$\widehat{\alpha}_j^{g(m)} = \frac{\sum_{i=1}^n \widehat{Z}_{ig}^{(m)} \sum_{b=1}^B \mathbb{1}_{\{J_1^{i,b}=j\}}}{B \sum_{i=1}^n \widehat{Z}_{ig}^{(m)}}, \qquad \widehat{\mathbf{P}}_{hj}^{g(m)} = \frac{\sum_{i=1}^n \widehat{Z}_{ig}^{(m)} \sum_{b=1}^B n_{hj}^{ib}}{\sum_{\ell=1}^D \sum_{i=1}^n \widehat{Z}_{ig}^{(m)} \sum_{b=1}^B n_{h\ell}^{ib}}, \tag{18}$$

where n_{hj}^{ib} is the number of $h \to j$ transitions for trajectory S_i^b.

As shown in Chen et al. (2016), the log likelihood is not bounded in the simple gamma mixture models and adding a penalty is required to ensure the consistency of maximum likelihood estimators. We add to function Q defined in (12) a penalty similar to the penalty given in Chen et al. (2016), which acts on the parameter a of the gamma distribution and is defined as follows

$$P_{en}\left(a_{\ell g}, \ell \in \mathcal{S}, g = 1, \ldots, G\right) = -\frac{1}{\sqrt{\sum_{i=1}^{n}\sum_{b=1}^{B} N\left(T_i^b\right)}} \sum_{g=1}^{G}\sum_{\ell \in \mathcal{S}} \left(a_{\ell g} + \ln a_{\ell g}\right) \tag{19}$$

where $a_{\ell g}$ is the parameter related to the sojourn time in state ℓ for mixture g. Note that the penalty does not explicitly take into account the parameter $\lambda_{\ell g}$ of the gamma distribution and depends on the sample size and the number of observed transitions.

Finally, the parameters of the sojourn time distributions can be estimated by maximizing the following expected partial penalized log-likelihood:

$$\sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{Z}_{ig}^{(m)} \sum_{b=1}^{B}\sum_{k=1}^{N(T_i^b)} \ln \phi_{J_k^{i,b}}\left(X_k^{i,b}\right) + P_{en}\left(a_{\ell g}, \ell \in \mathcal{S}, g = 1, \ldots, G\right), \tag{20}$$

with classical optimization procedures.

3.4 Choosing the starting point of the EM algorithm

A crucial issue for the EM algorithm is the choice of the value of the starting point $\boldsymbol{\theta}^{(0)}$. It is shown in Galmarini et al. (2017) that the time spent in each state provides interesting indicator to study TDS data. Thus, we have chosen to select the initial values of the EM algorithm by considering the Hartigan-Wong k-means algorithm (Hartigan and Wong, 1979) applied to the D dimensional vector of mean sojourn times in each state, with the Euclidean distance and $k = G$ clusters. A heuristic justification can be given by the fact that the identification of the mixture components seems to be easier for sojourn times. Indeed as seen in (5), the sojourn time distribution of any finite mixture of Markov renewal processes is identifiable when the family of sojourn time distributions is identifiable. Then, the method of moments is employed to get initial values of the parameters for the gamma distributions.

3.5 Selection of the number of mixture components

When the number of components G is unknown, an information criterion is used to select, in a data driven way, the number of mixture components. Such criterion relies on a compromise between the fit to the data and the complexity of the considered model, more complex models being less desirable. The Bayesian information criterion (BIC), which has good asymptotic properties (see Keribin (2000)), seems effective to select the number of components and is defined as follows,

$$\mathrm{BIC}(G) = q \ln\left(nB\right) - 2 \ln L\left(S_1^b, \ldots, S_n^b, b = 1, \ldots, B;\ \hat{\boldsymbol{\theta}}(G)\right), \tag{21}$$

where $\hat{\boldsymbol{\theta}}(G)$ is the estimation of $\boldsymbol{\theta}$ when a mixture of G components has been considered and $q = q(\boldsymbol{\theta}(G))$ is the number of free parameters to be estimated. In the general framework of G mixtures of semi-Markov chains considered in this work, we have

$$\begin{aligned} q &= G - 1 + G(D - 1 + D(D - 2) + D(D - 1)d) \\ &= GD(D - 1)(d + 1) - 1 \end{aligned} \tag{22}$$

whereas

$$q = GD(D + d - 1) - 1 \tag{23}$$

with $d = 2$, in the particular setting described in Section 3.3. The number of mixture components is selected by minimizing the BIC criterion,

$$\widehat{G} = \arg \min_{G \in \{1,\dots,G_{\max}\}} \text{BIC}(G) \tag{24}$$

where $G_{\max} \leq n$ is the maximum number of allowed mixture components. Other popular criteria are the Akaike information criterion (AIC), in which the term $q \ln nB$ in (21) which penalize the complexity of the model is replaced by $2q$ and the corrected AIC, denoted by AIC_c in which the term $q \ln nB$ is replaced by $2q + \frac{2q(q+1)}{nB-q-1}$.

4 A simulation study

A simulation study is conducted to evaluate the performances of the penalized and unpenalized EM algorithms under various mixture scenarios. We also measure the ability of the AIC and the BIC criteria to select the correct number of mixture components. Simulations are performed using R (R Development Core Team, 2018) and C++ with the Rcpp package (Eddelbuettel and François, 2011).

4.1 Simulation protocol and indicators of performance

In order to get realistic simulations, we simulated qualitative trajectories based on semi-Markov chains whose parameters were estimated on the real dataset presented in the Introduction of the paper. In that experiment, panelists evaluated three different chocolates, with a list of $D = 10$ attributes, the first one with 70% of cocoa, the second one with 70% of cocoa too but sweeter than the first one and a third one with 90% of cocoa (see (Visalli et al., 2016) for a more detailed presentation of the data). An experiment related to the tasting of the first chocolate is presented in Figure 1. The components of the renewal Markov process corresponding to each chocolate are estimated by maximum likelihood (see Lecuelle et al. (2018)), considering gamma distributed sojourn times with no anticipation effect, as in Section 3.3. We also evaluate the effect of the introduction of the penalty (19) on the accuracy of the estimates.

First, we consider a known number of components equal to two, with simulated sequences of 4 or 10 transitions. We study two cases of mixtures: the fist one with two

well separated sub populations (the chocolates with 70% and 90% of cocoa) and the second one with two populations with similar distributions (the two chocolates with 70% of cocoa).

Second, we assume that the number of components is unknown to evaluate the ability of the different information criteria presented in Section 3.5 to recover the true number of components in the population. We consider three different configurations. One with only one component (chocolate with 70% of chocolate), one with two well separated components (the chocolates with 70% and 90% of cocoa) and one with two similar components (the two chocolates with 70% of cocoa). The selection of the number of components is a difficult task and the information criteria do not always give good results with stochastic processes (see for example (Celeux and Durand, 2008)).

Thanks to our knowledge of these chocolates, we can assume that some transitions are not possible (occur with a probability zero). Taking this information into account, we can reduce the number of transition parameters to be estimated. We have 49 unknown probability transition parameters for the chocolate with 70% of cocoa, 62 unknown parameters when considering the two chocolates with 70% of cocoa and 69 unknown parameters when considering the chocolate with 70% of cocoa and the one with 90% of cocoa.

For simulating mixtures with $G = 2$ components, the number of individuals belonging to each component is randomly selected thanks to a binomial law $B(n, 0.5)$, meaning that $\pi_1 = \pi_2 = 1/2$. Then, for each type of chocolate, individual trajectories are simulated sequentially by selecting randomly the successive states and durations according to the estimated transition probabilities and dominance duration distributions. For each case, we simulated 500 datasets with sample of sizes $n = 60$, $n = 200$ and $n = 600$ and $B = 3$ replications.

In order to avoid computation issues when estimating the parameters related to the gamma distributions, the values of $\widehat{Z}_{ig}^{(m)}$ are rounded to 10^{-4} and the maximum likelihood estimation is only performed when there are more than 7 observations. Otherwise the gamma parameters are set to the values estimated on all the observations belonging to the corresponding mixture, independently of the state.

The number of maximal iterations of the EM algorithm is set to 100. *A posteriori* this was large enough because, for all the considered designs, convergence was achieved before 100 iterations.

To check if the transition matrices are well estimated, we consider the following relative error between the estimated transition matrices $\widehat{\mathbf{P}}^g$ and the transition matrices $\mathbf{P}^g$ used to generate the simulated data for component g:

$$\text{Err}(\mathbf{P}^g) = \frac{\left\|\mathbf{P}^g - \widehat{\mathbf{P}}^g\right\|_2^2}{\|\mathbf{P}^g\|_2^2}, \tag{25}$$

where $\|\mathbf{P}\|_2 = \text{tr}\,(\mathbf{P}'\mathbf{P})$ is the squared Frobenius norm of matrix $\mathbf{P}$. A similar error is computed for the initial probabilities:

$$\text{Err}(\boldsymbol{\alpha}^g) = \frac{\|\boldsymbol{\alpha}^g - \widehat{\boldsymbol{\alpha}}^g\|_2^2}{\|\boldsymbol{\alpha}^g\|_2^2}. \tag{26}$$

Table 1: Parameter estimation errors when considering unpenalized EM for two clusters with $n = 60, n = 200$ and $n = 600$ and with simulated sequences with 4 and 10 transitions and $B = 3$ repetitions. For each design, mean and standard deviation, in brackets, are computed considering 500 simulated datasets.

Parameters	Err($\boldsymbol{\alpha}^1$)	Err($\boldsymbol{\alpha}^2$)	Err($\mathbf{P}^1$)	Err($\mathbf{P}^2$)	Err(a)	Err(λ)	$\pi_1 = 0.5$
With 4 transitions							
Well separated components							
$n = 60$	.01(.01)	.01(.02)	.26(.12)	.18(.10)	.23(.75)	.40(.91)	.55(.14)
$n = 200$	<.01(<.01)	<.01(<.01)	.06(.04)	.04(.02)	.04(.06)	.08(.14)	.50(.04)
$n = 600$	<.01(<.01)	<.01(<.01)	.02(.01)	.01(<.01)	.01(.01)	.02(.02)	.50(.02)
Not well separated							
$n = 60$	.01(.01)	.03(.04)	.33(.13)	.47(.15)	.44(1.77)	.70(3.00)	.61(0.25)
$n = 200$	<.01(<.01)	.01(.03)	.10(.08)	.16(.16)	.29(2.18)	.38(2.93)	.49(.12)
$n = 600$	<.01(<.01)	<.01(<.01)	.02(.01)	.01(.03)	.03(.06)	.04(.12)	.50(.03)
With 10 transitions							
Well separated components							
$n = 60$	.01(.01)	.01(.01)	.08(.06)	.05(.04)	.07(.12)	.15(.21)	.50(.10)
$n = 200$	<.01(<.01)	<.01(<.01)	.02(.01)	.01(<.01)	.01(.01)	.02(.02)	.50(.04)
$n = 600$	<.01(<.01)	<.01(<.01)	.01(<.01)	<.01(<.01)	<.01(<.01)	.01(<.01)	.50(.02)
Not well separated							
$n = 60$	.01(.01)	.03(.07)	.09(.06)	.23(.19)	.18(.41)	.21(.85)	.62(.19)
$n = 200$	<.01(<.01)	<.01(<.01)	.02(.01)	.02(.06)	.03(.04)	.03(.04)	.53(.07)
$n = 600$	<.01(<.01)	<.01(<.01)	.01(<.01)	<.01(<.01)	.01(.01)	.01(.01)	.50(.02)

We also check if the estimated parameters of the sojourn time gamma distribution are well estimated by considering the following relative errors

$$\text{Err}(a) = \frac{\sum_{l=1}^{D}\sum_{g=1}^{G}((\widehat{a}_l^g) - a_l^g)^2}{\sum_{l=1}^{D}\sum_{g=1}^{G}(a_l^g)^2}, \qquad \text{Err}(\lambda) = \frac{\sum_{l=1}^{D}\sum_{g=1}^{G}((\widehat{\lambda}_l^g) - \lambda_l^g)^2}{\sum_{l=1}^{D}\sum_{g=1}^{G}(\lambda_l^g)^2}. \tag{27}$$

4.2 Results

Parameter estimation errors, evaluated with (25), (26) and (27), are given in Table 1 for the unpenalised version of the EM algorithm and in Table 2 when the penalized version of the EM algorithm described in (20) is employed to estimate the parameters. We note that the introduction of the penalty allows to improve the accuracy of the estimates, especially for small samples, few transitions, or with clusters with similar distribution of the semi-Markov processes. Without penalty, we observe larger mean errors for the estimated parameters of the sojourn time distribution and high values for the standard deviations of the errors. For example, when $n = 60$ with only 4 observed transitions and clusters that are not well separated, we obtain $Err(\lambda) = 0.70$ without penalty whereas this error is reduced to $Err(\lambda) = 0.22$ thanks to the introduction of the penalty. When

Table 2: Parameter estimation errors when considering penalized EM for two clusters with $n = 60, n = 200$ and $n = 600$ and with simulated sequences with 4 and 10 transitions and $B = 3$ repetitions. For each design, the mean and the standard deviation, in brackets, are computed considering 500 simulated datasets.

Parameters	Err($\boldsymbol{\alpha}^1$)	Err($\boldsymbol{\alpha}^2$)	Err($\mathbf{P}^1$)	Err($\mathbf{P}^2$)	Err(a)	Err(λ)	$\pi_1 = 0.5$
With 4 transitions							
Well separated components							
$n = 60$	.01(.01)	.01(.02)	.26(.13)	.18(.10)	.10(.07)	.24(.15)	.54(.15)
$n = 200$	<.01(<.01)	<.01(<.01)	.06(.04)	.04(.02)	.03(.03)	.06(.07)	.50(.05)
$n = 600$	<.01(<.01)	<.01(<.01)	.02(.01)	.01(<.01)	.01(<.01)	.01(.01)	.50(.02)
Not well separated							
$n = 60$	.01(.01)	.04(.08)	.32(.13)	.48(.16)	.11(.11)	.22(.42)	.61(.26)
$n = 200$	<.01(<.01)	.01(.02)	.10(.07)	.15(.16)	.09(.19)	.11(.22)	.50(.12)
$n = 600$	<.01(<.01)	<.01(<.01)	.02(.01)	.01(.05)	.03(.22)	.04(.30)	.50(.03)
With 10 transitions							
Well separated components							
$n = 60$	.01(.01)	.01(.01)	.08(.07)	.06(.05)	.06(.04)	.13(.11)	.49(.11)
$n = 200$	<.01(<.01)	<.01(<.01)	.01(.01)	.01(.01)	.01(.01)	.02(.03)	.50(.04)
$n = 600$	<.01(<.01)	<.01(<.01)	.01(<.01)	<.01(<.01)	<.01(<.01)	.01(<.01)	.50(.02)
Not well separated							
$n = 60$	.01(.01)	.02(.04)	.10(.06)	.20(.18)	.09(.14)	.10(.20)	.62(.18)
$n = 200$	<.01(<.01)	<.01(<.01)	.02(.01)	.01(.04)	.03(.02)	.03(.03)	.53(.07)
$n = 600$	<.01(<.01)	<.01(<.01)	.01(<.01)	<.01(<.01)	.01(<.01)	.01(<.01)	.50(.02)

Table 3: Correct classification rate for two clusters with well separated components and not well separated components with $n = 60, n = 200$ and $n = 600$ and with length of simulated sequences equal to 4 and 10 transitions. For each design, mean and standard deviation, in brackets, are computed from 500 simulated datasets.

	Well separated components			**Not well separated components**		
	$n = 60$	$n = 200$	$n = 600$	$n = 60$	$n = 200$	$n = 600$
With 4 transitions						
k-means	.85(.07)	.86(.05)	.86(.03)	.81(.09)	.78(.08)	.76(.06)
Mixture model	.92(.07)	.99(.02)	1(<.01)	.82(.09)	.93(.06)	.98(.02)
With 10 transitions						
k-means	.87(.07)	.89(.04)	.90(.02)	.83(.07)	.84(.05)	.85(.03)
Mixture model	.97(.05)	1(.01)	1(<.01)	.89(.09)	.97(.05)	1(<.01)

the sample size gets larger ($n = 200$ or $n = 600$) and the number of transitions is large both estimation procedures lead to similar results.

From now on, we will only consider estimates obtained with the penalized EM algorithm.

In our simulation context, we know for each trajectory which component of the mixture it belongs to and we can check if it has been assigned with the MAP criterion to the right component. The rate of correct classification is given in Table 3. We note that overall, the rate of well classified trajectories is high with values ranging from 0.83 to 1. Model-based clustering substantially improves the classification accuracy compared to k-means, except for the more difficult case with a small sample ($n = 60$), 4 transitions and clusters not well separated, were both approaches do not perform well.

We present in Table 4 the number of components selected by the BIC and the AIC. Whatever the number of individuals, the BIC and the AIC select the correct number of components when there is only one component. With two well separated clusters, the BIC and the AIC generally give good results, except for the case with 4 transitions and $n = 60$ where the BIC and, to a lesser degree, the AIC, select only one component rather than two.

When the two mixture components are not very different, the BIC and the AIC only provide effective criteria for selecting the number of components when the samples are large. The AIC often selects the same number of components as the BIC, but it sometimes selects too many components. For small samples and small number of transitions, the BIC criterion is more restrictive and tends to lead to an underestimation of the true number of components. Similar conclusions, in a different context, are found in Celeux and Durand (2008). The AIC_c can only be used with large samples because of the too large number of parameters of the model. It does not perform better than the BIC and the AIC in this simulation study and the corresponding results are not shown here.

Table 4: Choice of the number of components with one component, two well separated components and two not well separated components and 4 or 10 observed transitions. The number of clusters selected by the BIC and the AIC are shown for 500 simulated datasets.

Selected number of components	One component			Two components					
				Well separated			Not well separated		
n	60	200	600	60	200	600	60	200	600
With 4 transitions									
BIC									
1	500	500	500	500	5	0	500	491	0
2	0	0	0	0	493	500	0	9	494
3	0	0	0	0	2	0	0	0	6
AIC									
1	500	500	500	93	0	0	491	4	0
2	0	0	0	394	473	498	9	373	487
3	0	0	0	13	27	2	0	123	13
With 10 transitions									
BIC									
1	500	500	500	43	0	0	411	0	0
2	0	0	0	457	497	500	89	431	495
3	0	0	0	0	3	0	0	69	5
AIC									
1	500	500	500	0	0	0	27	0	0
2	0	0	0	407	491	498	245	318	495
3	0	0	0	93	9	2	228	182	15

Table 5: Values taken by the BIC, the AIC and the AIC_c for a number of clusters G ranging from 1 to 4 for the young and low fat Gouda cheese from the ESN dataset.

G	1	2	3	4
BIC	78804.94	78196.00	78325.83	78813.67
AIC	78194.71	76969.95	76483.96	76328.64
AIC_c	78237.92	77186.49	77132.14	78045.62

5 Clustering Temporal Dominance of Sensations for a Gouda cheese

We now study data resulting on an experiment of the European Sensory Network (ESN) aiming at measuring simultaneously perception and liking of Gouda cheeses (Thomas et al., 2017). A large panel of $n = 665$ consumers from 6 european countries tasted 4 Gouda cheeses with different ages and fat contents according to the Temporal Dominance of Sensations protocol. A list of $D = 10$ attributes was presented to the consumers on a computer screen. Panelists tasted $B = 3$ successive bites so there is 3 sequences corresponding to the 3 repetitions for each panelist and for each product. In this sample, the mean number of transitions within a sequence is equal to 4.1.

Our goal in this study is to perform a segmentation of the panel, and to describe, if there are any, the differences of perceptions for a product. We only present the results for a young and low fat Gouda cheese, whose perception by consumers is more complex.

The maximal number of iterations of the EM algorithm is set to 400 because we observed that the algorithm requires more iterations to converge with this dataset. This can be explained by a higher complexity of the model than in the simulation study because all transitions are possible with these products.

As shown in Table 5, all the information criteria approaches select at least two mixture components, showing the existence of different behaviors in the panel. The BIC suggests to consider two clusters and the AIC_c suggests to consider three clusters but both take really close values for two and three components. That is why, we will examine these two cases in the following. The AIC suggests to consider at least 4 clusters, but as it well known, it is a less parsimonious criterion than the BIC and the AIC_c which generally leads to consider a too large number of mixture components. With two components, the obtained clusters are respectively composed of 400 and 265 individuals, whereas with three components, the obtained clusters are respectively composed of 269, 183 and 213 individuals.

The estimated initial probabilities are shown in Table 6. As expected in sensory studies, most of the panelists choose a texture attribute (Dense hard or Tender) as first dominant attribute. With two components, the initial probabilities are really close with only some small differences. On the other hand, with three components, large differences are observed between clusters especially for the attributes Dense hard and Tender. In cluster one, most of the panelists chose Tender as first attribute whereas in cluster two,

Table 6: Estimated initial probabilities for the young and low fat Gouda cheese from the ESN dataset, considering two or three mixture components.

Cluster	*Estimates for the following attributes:*									
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender
With 2 components										
1	.03	.07	.44	.06	.03	.03	.03	.02	.02	.27
2	.02	.05	.39	.07	.05	.05	.03	.02	.01	.31
With 3 components										
1	.01	.08	.21	.07	.04	.06	.04	.01	.01	.44
2	.04	.06	.74	.03	.01	0	.03	.02	.02	.05
3	.02	.05	.41	.07	.05	.05	.02	.02	.01	.31

Table 7: Proximities between the estimated transition matrices when considering three components for the young and low fat Gouda cheese.

	$(\mathbf{P}^1, \mathbf{P}^2)$	$(\mathbf{P}^1, \mathbf{P}^3)$	$(\mathbf{P}^2, \mathbf{P}^3)$
Dist.	0.26	0.07	0.19

most of the panelists chose Dense hard. In cluster three, both Dense hard and Tender have a high probability to be chosen as first attribute.

Figure 3 presents the estimated gamma distributions of the sojourn times, with two components, for the attributes Cheese, Dense hard and Tender. Figure 4 presents the estimated sojourn time distributions when considering three components. We can note that for all clusters, there are only small differences between the estimated distributions of the different attributes. Then, we can observe that with two components, the estimated distributions are different between the two clusters, with higher probabilities for long durations in cluster one. With three components, the estimated distributions are really similar for the clusters one and two but are different from cluster three.

We now introduce a criterion that will be useful to measure a relative proximity between two transition matrices $\mathbf{P}^{g_1}$ and $\mathbf{P}^{g_2}$,

$$Dist(\mathbf{P}^{g_1}, \mathbf{P}^{g_2}) = \frac{||\mathbf{P}^{g_1} - \mathbf{P}^{g_2}||_2^2}{||\mathbf{P}^{g_1}||_2^2 + ||\mathbf{P}^{g_2}||_2^2}.$$

The proximities of the estimated transition matrices are given in Table 7 when considering three components. We note that cluster one and cluster two are well separated whereas the transition matrices for cluster one and cluster three are quite similar. With only two components, the two estimated transition matrices are very similar, with $Dist(\mathbf{P}^1, \mathbf{P}^2) = 0.05$, meaning that there is almost no difference between these two transition matrices.

To sum up, with two components, the main differences between the estimated distributions for cluster one and two come from the sojourn time distributions. If we consider

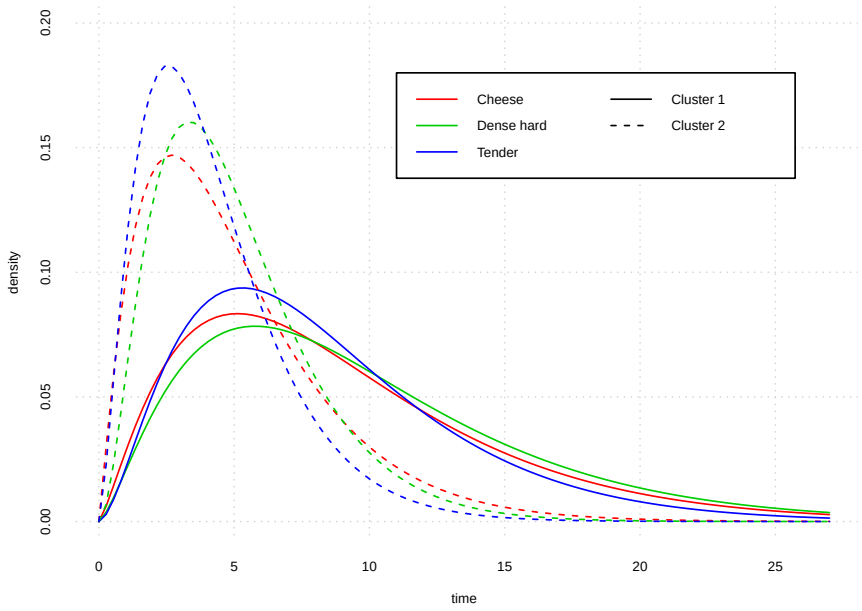

Figure 3: Estimated sojourn time distributions for the attributes Cheese, Dense hard and Tender when considering two components for the young and low fat Gouda cheese.

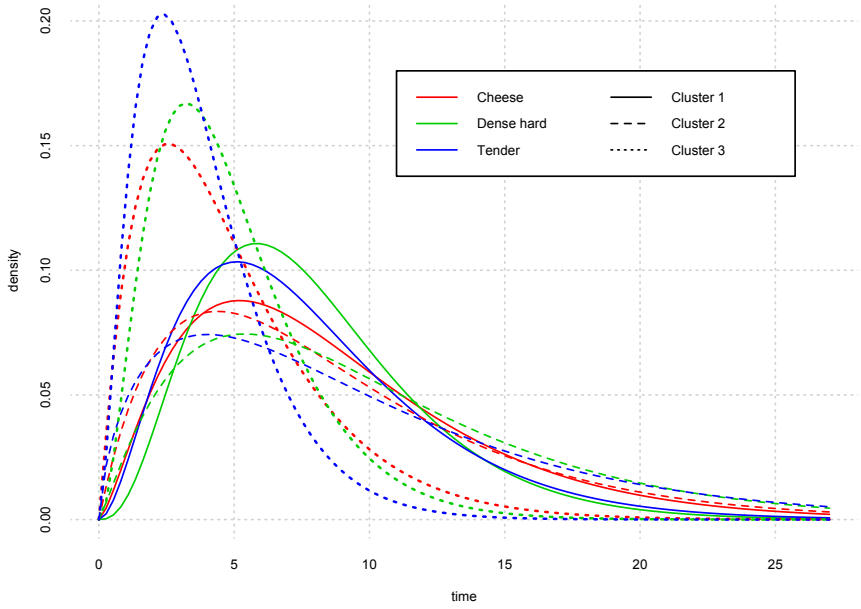

Figure 4: Estimated sojourn time distributions for the attributes Cheese, Dense hard and Tender when considering three components for the young and low fat Gouda cheese.

a mixture model with three components, clusters one and two are different according to their initial probabilities and to their transition matrices but have similar sojourn time distributions, as if one cluster in a two mixture component model has been split in two to get a three components mixture model. From a sensory perspective, the segmentation enables to model what is really perceived by the panelists instead of considering a mean panel overview, corresponding to the perception of none of the panelists. The observed differences between clusters can be explained both by real differences of perception and by differences of behavior with the TDS task. Such mixture models give the opportunity to further investigate on these new questions by for example examining the relation between perception and other variables such as age, sex or experience.

6 Concluding remarks

This research was motivated by the need of a segmentation method for temporal sensory data. For this purpose we have introduced a new mixture of semi-Markov chains which allows, thanks to a model-based clustering approach, to gather into homogeneous groups consumers having similar tasting perceptions. A penalized EM is introduced to estimate the parameters of the semi Markov chains and the mixture proportions. The evaluation of this estimation method on simulated data shows good performance, improving the segmentation obtained by the k-means algorithm, while providing much more information on individual behaviors. The results on real data show an interesting progress in TDS data analysis by offering the possibility of exhibiting different perceptions in a panel for a same product. The development of such segmentation approaches open new perspectives, both for understanding the perception mechanism and for studying how panelists used TDS and understand the TDS protocol.

The models presented in this paper may depend on a large number of parameters and so require to have large samples at hand to be estimated accurately. However, the real data analysis shows that only small differences seem to exist in a same cluster between the gamma distributions modeling sojourn times in the different states. If this hypothesis is verified, we could consider a more parsimonious model by estimating only one gamma distribution for all the states. As usual with unsupervised classification, choosing the number of clusters is a difficult task. The method used in this paper relies on information criteria and is not very effective for small samples. The BIC criterion seems to overestimate the model complexity whereas AIC has a tendency to select models with a too large complexity.

From a statistical perspective, this sensory modeling issue has given us the opportunity to study a new model for mixtures of qualitative trajectories which may have applications in many fields of science. The identifiability issue has been addressed under general conditions, considering parametric families of sojourn time distributions. From a methodological perspective, it also showed that introducing a penalty in the maximisation step of the EM algorithm improves the quality of the estimates. However some further investigations have to be done to determine whose penalty is the most effective. It would also be of great interest to check rigorously in a future work the consistency of

such penalized maximum likelihood approach in the context of mixtures of semi-Markov chains and to study the asymptotic distribution of the estimators. This would permit to build confidence intervals and to test statistical hypotheses.

References

Allman, E. S., C. Matias, and J. A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist. 37*(6A), 3099–3132.

Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*(3), 803–821.

Barbu, V. S. and N. Limnios (2008). *Semi-Markov chains and hidden semi-Markov models toward applications : their use in reliability and DNA analysis*. New York: Springer Science + Business Media.

Celeux, G. and J.-B. Durand (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist. 23*, 541–564.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statist. Sci. 32*(1), 47–63.

Chen, J., S. Li, and X. Tan (2016). Consistency of the penalized MLE for two-parameter gamma mixture models. *Sci. China Math. 59*(12), 2301–2318.

Delattre, M., V. Genon-Catalot, and A. Samson (2016). Mixtures of stochastic differential equations with random effects: Application to data clustering. *Journal of Statistical Planning and Inference 173*, 109–124.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological 39*(1), 1–38.

Eddelbuettel, D. and R. François (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40*(8), 1–18.

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

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer series in statistics. Springer.

Frydman, H. (2005). Estimation in the mixture of Markov chains moving with different speeds. *Journal of the American Statistical Association 100*(471), 1046–1053.

Galmarini, M. V., M. Visalli, and P. Schlich (2017). Advances in representation and analysis of mono and multi-intake temporal dominance of sensations data. *Food Quality and Preference 56*, 247–255.

Gupta, R., R. Kumar, and S. Vassilvitskii (2016). On mixtures of Markov chains. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29*, pp. 3441–3449. Curran Associates, Inc.

Hartigan, J. and M. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics 28*, 100–108.

Hort, J., S. Kemp, and T. Hollowood (2017). *Time-dependent measures of perception in sensory evaluation*. Wiley Blackwell.

Jaeger, S. R., J. Hort, C. Porcherot, G. Ares, S. Pecore, and H. J. H. MacFie (2017). Future directions in sensory and consumer science: Four perspectives and audience voting. *Food Quality and Preference 56*, 301–309.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya: The Indian J. of Statist., Serie A 62*, 49–66.

Lawlor, S. and M. G. Rabbat (2017). Time-varying mixtures of Markov chains: An application to road traffic modeling. *IEEE Transactions on Signal Processing 65*(12), 3152–3167.

Lecuelle, G., M. Visalli, H. Cardot, and P. Schlich (2018). Modeling temporal dominance of sensations with semi-Markov chains. *Food Quality and Preference 67*, 59–66.

Lévy, P. (1956). Processus semi-Markoviens. In *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, vol. III*, pp. 416–426. Erven P. Noordhoff N.V., Groningen; North-Holland Publishing Co., Amsterdam.

McLachlan, G. J. and T. Krishnan (2008). *The EM algorithm and extensions* (2nd ed.). New York: Wiley Series in Probability and Statistics.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley Series in Probability and Statistics.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification 33*(3), 331–373.

Melnykov, V. and R. Maitra (2010). Finite mixture models and model-based clustering. *Statistics Surveys 4*, 80–116.

Neilson, A. (1957). Time-intensity studies. *Drug & Cosmetic Industry 80*, 452–453.

Norris, J. R. (1998). *Markov chains*, Volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. Reprint of 1997 original.

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

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schlich, P. (2017). Temporal dominance of sensations (TDS): a new deal for temporal sensory analysis. *Current Opinion in Food Science 15*, 38–42.

Smith, W. L. (1955). Regenerative stochastic processes. *Proceedings of the Royal Society Series A 232*, 6–31.

Song, Y., A. Keromytis, and S. Stolfo (2009). Spectrogram: A mixture-of-Markov-chains model for anomaly detection in web traffic. In *Proceedings of the Network and Distributed System Security Symposium, NDSS*.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics 34*, 1265–1269.

Thomas, A., M. Chambault, L. Dreyfuss, C. C. Gilbert, A. Hegyi, S. Henneberg, A. Knippertz, E. Kostyra, S. Kreme, A. P. Silva, and P. Schlich (2017). Measuring temporal liking simultaneously to temporal dominance of sensations in several intakes. an application to gouda cheeses in 6 europeans countries. *Food Research International 99*, 426–434.

Titterington, D., A. Smith, and U. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Visalli, M., C. Lange, L. Mallet, S. Cordelle, and P. Schlich (2016). Should I use touch-screen tablets rather than computers and mice in TDS trials? *Food Quality and Preference 52*, 11–16.

Yakowitz, S. and J. Spragins (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics 39*, 209–214.