Electronic Journal of Statistics
Vol. 3 (2009) 1113–1132
ISSN: 1935-7524
DOI: 10.1214/09-EJS478

A class of models for multiple binary sequences under the hypothesis of Markov exchangeability

Davide Di Cecco

Viale Scalo S. Lorenzo 79, 00185 Roma
e-mail: davide.dicecco@gmail.com

Abstract: We discuss inference for multiple binary sequences under the hypothesis of Markov exchangeability. So far, the only kind of models for this purpose have been the mixtures of Markov chains. We present a new class of hierarchical models parameterized in terms of Bahadur/Lancaster's interactions, and compare it to the mixtures of Markov chains models.

AMS 2000 subject classifications: Primary 60G09; secondary 62M99.
Keywords and phrases: Markov exchangeability, multiple binary sequences, Bahadur/Lancaster's interaction.

Received September 2009.

1. Introduction

The concept of Markov exchangeability was suggested for the first time as a particular case of partial exchangeability by de Finetti in [6]. This particular case has peculiar importance and is sometimes simply referred to as partial exchangeability. Since then, the main aim of the authors working on this specific topic has been to find a theorem characterizing the discrete time mixtures of Markov chains processes. Various papers ([13, 14, 8, 34, 12]) outline different theoretical frameworks for extending de Finetti's representation theorem to Markov exchangeable (hereafter ME) processes, the main result being that a recurrent process is ME if, and only if, its law is a mixture of Markov chains. Besides this characterization problem, some particular ME processes have been studied in the context of certain random walks over graphs, namely the Edge Reinforced Random Walks (see [25] and references therein for a comprehensive survey), and in the Bayesian analysis of certain processes (see [24, 10]). The inferential analysis of ME processes has not received the same attention.

A finite sequence of r.v.s $(X_1, \ldots, X_n)$ taking values in a discrete state space I is said ME if its distribution assigns the same probability to all the I–valued sequences having the same starting state and the same number of transitions (i, j) for each couple of states in I. A process $\{X_n\}_{n\in\mathbb{N}}$ is said ME if the above holds for every n.

When we have to analyze a dataset consisting of multiple I–valued sequences, for example when we have a sequence of responses recorded for each unit, we may consider a hypothesis of Markov exchangeability. Rigorously, we should adopt it whenever, in a sequence of observations, we consider "the last outcome before any observation as a relevant attribute of the observation itself, and, once the observations are classified according to this attribute, time order becomes irrelevant" (see [12]). Even if we do not have such a precise idea on the data, we can adopt a hypothesis of Markov exchangeability for example whenever we deem it appropriate to use an homogeneous Markov chain, but this choice results in a poor fit to the data (possibly due to population's heterogeneity). Then we can look for a wider class of models maintaining the hypothesis of Markov exchangeability (a Markov chain is a particular ME process). Moreover, we can adopt it as a way to approximate a multivariate discrete distribution, reducing the complexity of the analysis. In fact, under Markov exchangeability, we a priori reduce the maximum number of free parameters the distribution can have (numerical details will be given on that). So, if one is interested, even if somehow informally, to the order of a set of I–valued variables, Markov exchangeability is a reasonable intermediate between considering the joint distribution in its greatest generality, and a simple Markov dependence (i.e., since I is discrete, a Markov chain).

To the author's knowledge, only Quintana and co–authors have explicitly considered this use of the hypothesis in several papers (see e.g. [26–28]). Other papers analyze the practical use of discrete time mixtures of Markov chains models for the analysis of multiple categorical sequences (e.g. [15, 17, 16]). Specifically, they analyze various kind of finite mixtures of Markov chains and their use in (model–based) cluster analysis, but they do not even mention the concept of Markov exchangeability.

This paper tries to fill the gap, analyzing the practical use of this hypothesis, and proposing an alternative to the mixtures of Markov chains models.

The need of an alternative to the mixture models is explained by the following argument: de Finetti's representation theorem asserts that an infinite sequence of r.v.s is exchangeable if, and only if, its law is a mixture of laws of i.i.d. variables. Not all the exchangeable finite sequences are initial segments of longer exchangeable sequences, i.e., as is said, not all of them are "extendible". Then, it may be the case that a mixture of i.i.d. model is not suitable to analyze the data at hand under the exchangeability assumption. As a simple example of this, in a mixture of i.i.d. r.v.s, the correlation among the variables is necessarily nonnegative, while it is not so in the not extendible case. The question of extendibility has been studied mainly for binary exchangeable r.v.s (e.g. [9, 3, 33]), and the concept has been extended to ME r.v.s in [35, 36], so that we can say that a finite ME sequence is not necessarily the initial segment of a mixture of Markov chains process.

We will consider the case $I = \{0, 1\}$, as repeated binary variables often arise in practice, and we present a new class of hierarchical models for ME binary data. Such a class is presented as a reparameterization of the joint distribution of n ME r.v.s in terms of the "Bahadur/Lancaster's interactions". These interactions, also called "additive interactions", were first introduced in [2] and [21] and define the "additive models" which constitute an alternative to the loglinear models for the analysis of categorical variables. In fact, this class of additive models for the ME case is large enough to include all the distributions of the ME sequences of a fixed length disregarding their extendibility under a simple additional assumption, and nearly all the others.

The paper is structured as follows: In Section 2 we give some definitions and insights in ME distributions and mixtures of Markov chains, and we present a first simple parameterization of a ME distribution. In Section 3 we define a second parameterization which serves as a necessary intermediate step in order to construct the Bahadur/Lancaster's interaction parameters, and an assumption we should adopt to successfully accomplish that construction. In Section 4 we briefly introduce the additive models in general, then present the additive models for ME binary data. An application of the models presented concludes the paper.

2. Some definitions

Consider an I–valued sequence $(x_1, \dots, x_n)$. Define its transition counts $n_{\,i,j}$ for all i, j in I as

$$n_{\,i,j} = \sum_{k=1}^{n-1} \mathbb{1}_{(i,j)}(x_k, x_{k+1})$$

Arrange them in a matrix $N = \{n_{\,i,j}\}_{i,j}$. Then, we will say that $(X_1, \dots, X_n)$ is ME (or n–ME when we need to highlight the length of the sequence), if its joint n–variate distribution assigns the same probability to all the $(x_1, \dots, x_n)$ in I^n having the same value of the first step x_1, and the same transition count matrix N. That is, x_1 and N together are a sufficient statistic, the probability of having any sequence starting in x_1 and consistent with N depends only on x_1 and N, and we will denote it $p_{\,x_1,N}$. Denote the set of all the distinct transition count matrices of all the I–valued sequences of n steps starting in x_1 as $\Phi(x_1, n)$. For what we have said, an n–ME distribution is completely defined by the probabilities $\{p_{\,x_1,N}\}$ for x_1 ranging in I and N ranging in $\Phi(x_1, n)$, and that constitutes a first simple parameterization of an n–ME distribution.

When $I = \{0, 1\}$, we deal with 2×2 transition count matrices of the kind:

$$N = \begin{pmatrix} n_{\,0,0} & n_{\,0,1} \\ n_{\,1,0} & n_{\,1,1} \end{pmatrix}$$

Two cases are possible: $n_{\,0,1}$ and $n_{\,1,0}$ are equal or differ by one. In the first case, the sequences necessarily start and end at the same state; in the second, in different states. So, if we know x_1 and N, we also know x_n. Let us denote as $\Phi_0(0, n)$ the subset of $\Phi(0, n)$ of the matrices having $n_{\,0,1} = n_{\,1,0}$. The corresponding sequences all start and end in 0. Denote as $\Phi_1(0, n)$ the subset of $\Phi(0, n)$ of the matrices having $n_{\,0,1} = n_{\,1,0} + 1$. The corresponding sequences all start in 0 and end in 1. We have $\Phi_0(0, n) \cup \Phi_1(0, n) = \Phi(0, n)$. Symmetrically, for the sequences starting in 1, we define

$$\Phi_0(1, n) = \{N \in \Phi(1, n) \,:\, n_{\,1,0} = n_{\,0,1} + 1\}$$
$$\Phi_1(1, n) = \{N \in \Phi(1, n) \,:\, n_{\,1,0} = n_{\,0,1}\}$$

such that $\Phi_0(1, n) \cup \Phi_1(1, n) = \Phi(1, n)$.

The number of probabilities defining an n–ME distribution is equal to the number of possible different transition count matrices for each fixed starting state. We have (for a proof see [7])

$$|\Phi(1,n)| = |\Phi(0,n)| = \binom{n}{2} + 1$$

So, in general an n–ME binary distribution is defined by the $2\binom{n}{2}+2$ probabilities $\{p_{0,N}\}_{N\in\Phi(0,n)}$ and $\{p_{1,N}\}_{N\in\Phi(1,n)}$. We will also use the symbols $p_{x_1, \left(\begin{smallmatrix} n_{0,0} & n_{0,1} \\ n_{1,0} & n_{1,1}\end{smallmatrix}\right)}$.

Let $\mathbf{N}$ be the transition count matrix intended as a r.v. Denote as $w_{x_1,N}$ the probability of $\big\{(X_1 = x_1) \cap (\mathbf{N} = N)\big\}$. Then any n–ME binary distribution is as well defined by the probabilities $\{w_{0,N}\}_{N\in\Phi(0,n)}$ and $\{w_{1,N}\}_{N\in\Phi(1,n)}$. The relation with the previous parameterization is clear: the parameter $w_{x_1,N}$ is the probability of having any sequence in I^n consistent with $\big\{(X_1 = x_1)\cap(\mathbf{N} = N)\big\}$, i.e. starting in x_1 and having the transition count N, and all those sequences have the same probability $p_{x_1,N}$. The number of sequences such defined has been first computed by Whittle in [32]. When $I = \{0,1\}$ and $N = \left(\begin{smallmatrix} n_{0,0} & n_{0,1} \\ n_{1,0} & n_{1,1}\end{smallmatrix}\right)$, the corresponding number is $\binom{n_0^+}{n_{0,0}}\binom{n_1^+-1}{n_{1,1}}$ if $x_n = 0$ and $\binom{n_0^+-1}{n_{0,0}}\binom{n_1^+}{n_{1,1}}$ if $x_n = 1$ where $n_0^+ = n_{0,0} + n_{0,1}$ and $n_1^+ = n_{1,0} + n_{1,1}$. Then we have

$$w_{x_1,N} = \begin{cases} \binom{n_0^+}{n_{0,0}}\binom{n_1^+-1}{n_{1,1}}\, p_{x_1,N} & \text{when } (x_1,N) \text{ lead to } x_n = 0 \\ \binom{n_0^+-1}{n_{0,0}}\binom{n_1^+}{n_{1,1}}\, p_{x_1,N} & \text{when } (x_1,N) \text{ lead to } x_n = 1 \end{cases} \tag{1}$$

If we do not add any assumption, the $\{w_{x_1,N}\}$ are subject to the only restriction:

$$\sum_{x_1\in I}\ \sum_{N\in\Phi(x_1,n)} w_{x_1,N} = 1$$

Then we have $2\binom{n}{2} + 1$ free parameters, and that is the maximum number of identifiable free parameters for an n–ME distribution. Any parameterization of an n–ME distribution without further assumptions would be a one–to–one transform of the $\{w_{x_1,N}\}$ (or equivalently of the $\{p_{x_1,N}\}$), and so would have that number of free parameters. Then, in a fully parametric approach, we will say that a model for an n–ME sequence is saturated if it has $2\binom{n}{2} + 1$ free parameters. That result allows us to appreciate the usefulness of a hypothesis of Markov exchangeability in terms of reduction of complexity: if we do not make any assumption on the joint distribution of n binary r.v.s, we would have $2^n - 1$ free parameters. Under the only assumption of Markov exchangeability, we a priori reduce that number to about n^2. In case of simple exchangeability it would be n, but we would have lost any information about the order of the variables.

2.1. Mixtures of Markov chains

We say that an I–valued process $X = \{X_n\}_{n\in\mathbb{N}}$ is ME if $(X_1,\dots,X_n)$ is ME for every n. Diaconis and Freedman in [8] demonstrated that a recurrent process

($X_1 = X_n$ i.o.) is ME if, and only if, its law is a mixture of Markov chains. That is, let $\mathcal{P}$ be the space of all the stochastic matrices $\Theta = \{\theta_{\,i,j}\}_{i,j}$ on $I \times I$. Then there exists a mixing measure ν on the Borel sets of $I \times \mathcal{P}$ such that

$$P(X_1 = x_1, \dots, X_n = x_n) = \int_{\mathcal{P}} \prod_{i=1}^{n-1} \theta_{x_i, x_{i+1}}\, \nu(x_1, d\Theta)$$

Let $\Gamma_i(k)$ be the step of the process at which the state i occurs for the k–th time. Let $V_i(k)$ be the k–th successor of the state i, i.e. the variable immediately subsequent the k–th occurrence of i: $V_i(k) = X_{\Gamma_i(k)+1}$, and let $v_i(k)$ be the corresponding observed value. Originally de Finetti hinted at the possibility to characterize X as a mixture of Markov chains by the exchangeability of all the subprocesses $\{V_i(k)\}_k$, $i \in I$. Much later in [12] it has been demonstrated that the idea of de Finetti and the characterization of Diaconis and Freedman coincide in case of recurrent processes. In the following we will use the fact that in a ME process the $\{V_i(k)\}_k$ are exchangeable.

When $I = \{0, 1\}$, the only non–recurrent ME processes are negligible degenerate cases (see [8]). In a recurrent binary ME process there are two exchangeable subprocesses $\{V_0(k)\}_k$ and $\{V_1(r)\}_r$, and there exists a measure ν defined on the Borel sets of $I \times [0, 1]^2$ determining the joint distribution of $(X_1, \theta_{\,0,0}, \theta_{\,1,1})$, such that

$$\begin{aligned} p_{\,x_1,N} &= p_{\,x_1,\left(\begin{smallmatrix} n_{\,0,0} & n_{\,0,1} \\ n_{\,1,0} & n_{\,1,1}\end{smallmatrix}\right)} \\ &= \int_0^1\!\!\int_0^1 \theta_{\,0,0}^{\,n_{\,0,0}} \,(1-\theta_{\,0,0})^{n_{\,0,1}}\ \theta_{\,1,1}^{\,n_{\,1,1}}(1-\theta_{\,1,1})^{n_{\,1,0}}\ \nu\,(x_1, d\theta_{\,0,0}, d\theta_{\,1,1}) \end{aligned} \quad (2)$$

If ν factorizes: $\nu(x_1, \theta_{\,0,0}, \theta_{\,1,1}) = \nu_1(x_1, \theta_{\,0,0})\, \nu_2(x_1, \theta_{\,1,1})$, for $x_1 = 0, 1$, the two exchangeable subprocesses $\{V_0(k)\}_k$ and $\{V_1(r)\}_r$ are independent. If ν is concentrated on the diagonal set $\{(\theta, 1-\theta) \mid \theta \in [0, 1]\}$, we obtain an exchangeable binary process. If ν is concentrated in a single point of the unit square we obtain an ordinary Markov chain. Furthermore, if this point belongs to the diagonal set above, the resulting is an i.i.d. process.

We will consider two kinds of mixtures of Markov chains models. In the first model, introduced in [27], X_1, $\theta_{\,0,0}$ and $\theta_{\,1,1}$ are hypothesized to be independent, X_1 is modelled separately, and both $\theta_{\,0,0}$ and $\theta_{\,1,1}$ have a Beta mixing distribution. That is, $d\nu\,(\theta_{\,0,0}, \theta_{\,1,1})$ can be written as

$$Beta(\theta_{\,0,0}\,;\ \alpha_0, \beta_0)\ Beta(\theta_{\,1,1}\,;\ \alpha_1, \beta_1)\ d\mu(\theta_{\,0,0}, \theta_{\,1,1})$$

where $Beta(\,\cdot\,; \alpha, \beta)$ is the Beta density of parameters α and β, and $\mu(\theta_{\,0,0}, \theta_{\,1,1})$ is the Lebesgue measure on $[0, 1]^2$. The above product of two independent Beta is sometimes called Matrix Beta distribution. So, we will call the resulting mixture of Markov chains Matrix Beta Mixture (MBM). The MBM model is defined by 5 free parameters: $\alpha_0, \beta_0, \alpha_1, \beta_1$ and $P(X_1 = 1) = q_1$.

The second class of mixtures of Markov chains models we will consider are the finite mixture models (see [17, 28]), where the mixing distribution is conceived as a discrete distribution supported only on a finite number of points. A binary simple Markov chain is completely defined by three parameters: the probability of transition $(0,0)$, $\theta_{0,0}$, the probability of transition $(1,1)$, $\theta_{1,1}$, and by $P(X_1 = 1) = q_1$. Then if the discrete mixing distribution has, say, d support points, it assigns masses λ_h, $h = 1,\dots,d$, $\sum_{h=1}^{d} \lambda_h = 1$ to d points $\big(q_1(h),\ \theta_{0,0}(h),\ \theta_{1,1}(h)\big)$ in the parameter space of a Markov chain. The Markov chains having these parameter's values are called the component Markov chains and d is the number of components. The λ_h are called the mixing weights. Thus, we have to estimate $(d-1)$ free mixture weights and three parameters $q_1(h)$, $\theta_{0,0}(h)$, $\theta_{1,1}(h)$ for each component Markov chain, that is, a total of $4d-1$ independent parameters.

When dealing with mixtures of discrete components distributions, an identifiability problem may arise. That question has been studied and solved in the case of finite mixtures of Binomials, calculating the maximum number of components the mixture can have, that guarantee the identifiability of all the parameters (see, for example, [22] for a geometric approach). In our case, we have already calculated the maximum number of identifiable parameters $2\binom{n}{2}+1$, then $4d-1$ could not exceed that number and we have:

$$d \leq \frac{\binom{n}{2}+1}{2}$$

In [28] the authors consider a mixing distribution over d points $\big(\theta_{0,0}(h), \theta_{1,1}(h)\big)$, $h = 1,\dots,d$, i.e., X_1 is analyzed separately and the total number of free parameters is $3d$.

3. A first reparameterization

In an n–ME sequence $(X_1,\dots,X_n)$, the first k elements $(X_1,\dots,X_k)$, $k < n$, constitute a k–ME sequence, and we can obtain all the probabilities of the kind $\{p_{x_1,K}\}_{K\in\Phi(x_1,k)}$ starting from the $\{p_{x_1,N}\}_{N\in\Phi(x_1,n)}$. Let $K = \left(\begin{smallmatrix} k_{0,0} & k_{0,1} \\ k_{1,0} & k_{1,1}\end{smallmatrix}\right)$ be the transition count matrix up to the first k steps, i.e. $\sum_{i,j\in I} k_{i,j} = k-1$ and let $k_{0,0} + k_{0,1} = k_0^+$ and $k_{1,0} + k_{1,1} = k_1^+$. Then we have (for a proof see [7])

$$\begin{aligned} p_{0,K} = & \sum_{N\in\Phi_0(0,n)} \binom{n_0^+ - k_0^+}{n_{0,0} - k_{0,0}} \binom{n_1^+ - k_1^+ - 1}{n_{1,1} - k_{1,1}} \; p_{0,N} + \\ & + \sum_{N\in\Phi_1(0,n)} \binom{n_0^+ - k_0^+ - 1}{n_{0,0} - k_{0,0}} \binom{n_1^+ - k_1^+}{n_{1,1} - k_{1,1}} \; p_{0,N} \qquad (3) \end{aligned}$$

where the sums should be restricted over those matrices N in $\Phi(0,n)$ having $n_{i,j} \geq k_{i,j}$, for all i, j in I. A similar formula holds for the sequences starting in 1. Consider now the probability $p_{0,\left(\begin{smallmatrix} k & 1 \\ 0 & r\end{smallmatrix}\right)}$ of having the sequence of $r+k+2$ steps starting in 0 with k transitions $(0,0)$, a single transition $(0,1)$ and ending with r transitions $(1,1)$, and denote it $w_{\,0,k,r}$. Applying the above formula we have

$$w_{\,0,k,r} = \sum_{N\in\Phi_0(0,n)} \binom{n_0^+ - k - 1}{n_{\,0,0} - k}\binom{n_1^+ - r - 1}{n_{\,1,1} - r}\; p_{\,0,N} + \\ + \sum_{N\in\Phi_1(0,n)} \binom{n_0^+ - k - 2}{n_{\,0,0} - k}\binom{n_1^+ - r}{n_{\,1,1} - r}\; p_{\,0,N} \qquad (4)$$

We set $w_{\,0,n-1,0} = p_{\,0,\left(\begin{smallmatrix} n-1 & 0 \\ 0 & 0\end{smallmatrix}\right)}$.

Introduce the operators Δ_0 and Δ_1 such that:

$$\Delta_0\left(w_{\,0,k,r}\right) = w_{\,0,k+1,r} - w_{\,0,k,r} \qquad \text{and} \qquad \Delta_1\left(w_{\,0,k,r}\right) = w_{\,0,k,r+1} - w_{\,0,k,r}$$

then the inverse formula of (4), defining the $\{p_{\,0,N}\}$ in terms of the $\{w_{\,0,k,r}\}$, is the following (see [7]):

$$\begin{aligned} p_{\,0,N} &= (-1)^{n_{\,0,1}-1+n_{\,1,0}}\Delta_0^{n_{\,0,1}-1}\,\Delta_1^{n_{\,1,0}}\left(w_{\,0,n_{\,0,0},n_{\,1,1}}\right) \\ &= \sum_{i=0}^{n_{\,0,1}-1}\sum_{j=0}^{n_{\,1,0}}(-1)^{\,i+j}\binom{n_{\,0,1}-1}{i}\binom{n_{\,1,0}}{j}\,w_{\,0,n_{\,0,0}+i,n_{\,1,1}+j} \end{aligned} \qquad (5)$$

In an n–ME sequence the probabilities $\{w_{\,0,k,r}\}$ are well defined for every couple of nonnegative integers (k,r) such that $0 \le k+r \le n-2$ together with the case $w_{\,0,n-1,0}$. The two formulas above assure that the set $\{w_{\,0,k,r}\}$ such defined constitutes a saturated parameterization of an n–ME distribution starting in 0. It is easily seen that the number of parameters defined is $\binom{n}{2}+1$. For the sequences starting in 1, we introduce the parameters $\{w_{\,1,k,r}\}$, defined as the probabilities of having the sequence starting in 1 with r transitions $(1,1)$, a single transition $(1,0)$ and ending with k transitions $(0,0)$. Formulas analogous to (4) and (5) define their one–to–one relation with the $\{p_{\,1,N}\}_{N\in\Phi(1,N)}$.

Let $Y_{i,j}(k)$ be the indicator function of the event $\big\{$the k–th successor of i is $j\big\}$ considered as a r.v. and let $y_{\,i,j}(k)$ be the corresponding observed value. That is:

$$\mathbb{1}_j\big(V_i(k)\big) = Y_{i,j}(k) \qquad\qquad \mathbb{1}_j\big(v_i(k)\big) = y_{\,i,j}(k) \qquad \forall\, i,j \in I$$

Note that the exchangeability of $\{V_0(k)\}_k$ and $\{V_1(r)\}_r$ implies the exchangeability of the $\{Y_{i,j}(k)\}_k$ for each fixed couple (i,j). We have

$$\begin{aligned} w_{\,0,k,r} &= E\Big[(1-X_1)\cdot Y_{0,0}(1)\cdots Y_{0,0}(k)\cdot\big(1-Y_{0,0}(k+1)\big)\cdot Y_{1,1}(1)\cdots Y_{1,1}(r)\Big] \\ w_{\,1,k,r} &= E\Big[X_1\cdot Y_{1,1}(1)\cdots Y_{1,1}(r)\cdot\big(1-Y_{1,1}(r+1)\big)\cdot Y_{0,0}(1)\cdots Y_{0,0}(k)\Big] \end{aligned}$$

In the particular case when $(X_1,\ldots,X_n)$ is the initial segment of a mixture of Markov chains process, that parameters are restricted to satisfy:

$$\begin{aligned}
w_{\,0,k,r} &= \int_0^1\!\!\int_0^1 (\theta_{\,0,0})^k(1-\theta_{\,0,0})(\theta_{\,1,1})^r\nu(0,d\,\theta_{\,0,0},d\,\theta_{\,1,1})\\
&= E_\nu\Big[(1-X_1)\ (\theta_{\,0,0})^k\ (1-\theta_{\,0,0})\ (\theta_{\,1,1})^r\Big]\\
w_{\,1,k,r} &= \int_0^1\!\!\int_0^1 (\theta_{\,0,0})^k(1-\theta_{\,1,1})(\theta_{\,1,1})^r\nu(1,d\,\theta_{\,0,0},d\,\theta_{\,1,1})\\
&= E_\nu\Big[X_1\ (\theta_{\,0,0})^k\ (1-\theta_{\,1,1})\ (\theta_{\,1,1})^r\Big]
\end{aligned}$$

where E_ν indicates the expectation w.r.t. the mixing measure $\nu(X_1,\theta_{\,0,0},\theta_{\,1,1})$.

Consider now the parameters $m_{\,i,k,r}$ defined as

$$m_{\,i,k,r} = E\big[\mathbb{1}_{\,i}\,(X_1)\cdot Y_{0,0}(1)\cdots Y_{0,0}(k)\cdot Y_{1,1}(1)\cdots Y_{1,1}(r)\big] \qquad i\in\{0,1\}$$

We have:

$$w_{\,0,k,r} = m_{\,0,k,r} - m_{\,0,k+1,r} \qquad \text{and} \qquad w_{\,1,k,r} = m_{\,1,k,r} - m_{\,1,k,r+1} \tag{6}$$

Then, once you know the $\{m_{\,i,k,r}\}$ defined for i in $\{0,1\}$ and every couple (k,r) such that $0\le k+r\le n-1$, excluding the cases $m_{\,0,0,n-1}$ and $m_{\,1,n-1,0}$, you can define an n–ME distribution. On the converse, it is easily seen that in an n–ME distribution it is not possible to single out all the values $\{m_{\,i,k,r}\}$ such defined without adding some restrictions, i.e., you cannot write them as a function of the $\{p_{\,i,N}\}$ and they are not identifiable. To realize that, one can simply note that their number is $2\binom{n+1}{2}-2$ which is greater than the maximum number of identifiable parameters for an n–ME distribution.

In order to reduce the total number of independent parameters and to make further constructions, we present an assumption which, though arbitrary, is reasonable and not particularly restrictive: we will call it "Independence Assumption", hereafter Ind.Ass. 1:

Ind.Ass. 1. *X_1 and $\boldsymbol{N}$ are independent.*

In a mixture of Markov chains model that assumption corresponds to

$$\nu(X_1,\theta_{\,0,0},\theta_{\,1,1}) = \nu_1(X_1)\,\nu_2(\theta_{\,0,0},\theta_{\,1,1})$$

Denote $P(X_1=i) = \displaystyle\sum_{N\in\Phi(i,n)} w_{\,i,N}$ as q_i and define

$$m_{\,k,r} = E\left[Y_{0,0}(1)\cdots Y_{0,0}(k)\cdot Y_{1,1}(1)\cdots Y_{1,1}(r)\right]$$

If we adopt Ind.Ass. 1, we can bypass the identifiability problem we have mentioned above. In fact, we have

$$m_{\,i,k,r} = q_i\ m_{\,k,r} \tag{7}$$

so, by (6), the $\{m_{\,k,r}\}$ defined for every (k,r) such that $0\le k+r\le n-1$ suffice to define an n–ME distribution under Ind.Ass. 1. On the converse, it is possible to write (with a recursive formula) all the probabilities $\{m_{k,r}\}$ such defined in terms of the $\{w_{0,k,r}\}$ and $\{w_{1,k,r}\}$. In fact we have

$$\frac{w_{0,k,r}}{q_0} = m_{k,r} - m_{k+1,r} \qquad \text{and} \qquad \frac{p_{1,\left(\begin{smallmatrix}0 & 0\\ 0 & r\end{smallmatrix}\right)}}{q_1} = m_{0,r}$$

Then we have $m_{1,r} = m_{0,r} - \frac{w_{0,0,r}}{q_0}$, and in general, by recurrence,

$$m_{k,r} = m_{k-1,r} - \frac{w_{0,k-1,r}}{q_0} \tag{8}$$

So, we can parameterize any n–ME distribution satisfying Ind.Ass. 1 in terms of the $\{m_{k,r}\}$, for $k + r \leq n - 1$, together with q_1.

In case of a mixture of Markov chains (2), the parameters $m_{k,r}$ are constrained to be the mixed moments of the mixing measure $\nu(\theta_{0,0}, \theta_{1,1})$:

$$m_{k,r} = E_\nu \left[\theta_{0,0}^{\,k}\, \theta_{1,1}^{\,r}\right]$$

Another case of interest is when the two exchangeable subprocesses forming the ME process are independent. We will refer to that condition as:

Ind.Ass. 2. *$\{Y_{0,0}(k)\}_k$ and $\{Y_{1,1}(r)\}_r$ are mutually independent sets.*

Ind.Ass. 1 and 2 together hold if and only if $m_{k,r} = m_{k,0}\, m_{0,r}$. The MBM and the simple Markov chain fall within this case. In particular, we have a simple Markov chain of parameters q_i, $\theta_{0,0}$ and $\theta_{1,1}$, if, and only if

$$m_{i,k,r} = q_i\, (\theta_{0,0})^k (\theta_{1,1})^r \qquad \text{and} \qquad m_{k,r} = (\theta_{0,0})^k (\theta_{1,1})^r \tag{9}$$

4. The additive models

A typical approach to the analysis of multivariate categorical data is that of defining some kind of interactions between the variables, by means of which we decompose their joint distribution in a hierarchical model. We can start from the saturated model, that is, when all the identifiable interaction parameters are included, to analyze the dependence structure. Then, setting equal to zero some interaction parameters, we can construct all the reduced models, and choose a suitable, parsimonious one, tuning our target of goodness of fit. Note that it is required that the interaction parameters can consistently assume zero values. There are two main approaches to interaction: the so called multiplicative and the additive approach. For a comparison of the two see [4, 5, 18]. The most commonly used is the multiplicative approach and the respective models, namely the loglinear models. The additive definition was introduced in [2] and defined for general real variables in [21] and hence is sometimes called Bahadur/Lancaster's interaction. It has been studied in few isolated papers ([37, 31, 30]). The respective models for categorical variables (we will call them additive models) have been rarely utilized ([23, 11]).

When dealing with $\{0,1\}$–valued variables, the additive interaction of order k among the variables $(X_1,\ldots,X_k)$ is

$$E[(X_1 - E[X_1])\cdots(X_k - E[X_k])]$$

It can be viewed as a generalization of the concept of covariance between two variables, so we will denote it $Cov[X_1,\ldots,X_k]$.

When data consist of sequences of unequal sizes, we should specify model parameters in such a way that they have a consistent interpretation, whatever the sequence size, i.e. parameters with meanings that are invariant across different marginal distributions. That is, the joint distribution of $(X_1,\ldots,X_k)$, $k < n$, should be defined by the same parameters, or a subset of the parameters, defining the joint distribution of $(X_1,\ldots,X_n)$. Sometimes, those distributions defined by parameters invariant under marginalization are said "reproducible". For some detail on the concept of reproducibility see [29, 18, 11].

The main advantage of the additive models with respect to the loglinear models is that the former are reproducible. On the converse, consider the following example from [18]: Let (X_1,X_2,X_3) be binary variables and denote $P(X_1 = i, X_2 = j, X_3 = h)$ as $p_{\,i,j,h}$, $i,j,h \in \{0,1\}$. Under the saturated loglinear model we have

$$\lg p_{\,i,j,h} = u + u_i(1) + u_j(2) + u_h(3) + u_{i,j}(1,2) + u_{i,h}(1,3) + u_{j,h}(2,3) + u_{i,j,h}(1,2,3)$$

where for example $u_{j,h}(2,3)$ is the interaction parameter resulting from the joint occurrences of state j in X_2 and state h in X_3. The logarithm of its marginal distribution for (X_1,X_2) is

$$\lg p_{\,i,j}(1,2) = \tilde{u} + \tilde{u}_i(1) + \tilde{u}_j(2) + \tilde{u}_{i,j}(1,2)$$

but $\tilde{u}_i(1) \neq u_i(1)$, $\tilde{u}_j(2) \neq u_j(2)$ and $\tilde{u}_{i,j}(1,2) \neq u_{i,j}(1,2)$. So loglinear models are not reproducible, and when data consist of multiple sequences of unequal sizes, we should prefer an additive model.

4.1. Additive models for Markov exchangeable data

The additive models for general r.v.s are a parameterization of the joint distribution of a set of variables. That parameterization is readily adaptable to the case of exchangeable binary variables. In fact, in that case the only difference is that things simplifies, as all the interactions of a same order are equal, i.e. they depend only on the number k of variables involved: $Cov[X_{i_1},\ldots,X_{i_k}] = Cov_k$, $\forall\ i_1,\ldots,i_k$. For an application of the additive models in the case of binary exchangeable data see [1] and [20].

In our case of ME variables, we do not define the additive interactions between the $X_1,\ldots,X_n$, but between the r.v.s $\{Y_{0,0}(k)\}_k$ and $\{Y_{1,1}(r)\}_r$, and we need an additional assumption to define the parameterization, namely Ind.Ass. 1.

Say data consist of m binary sequences $\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$. For each observed sequence $\mathbf{x}_l = (x_{l,1},\ldots,x_{l,n_l})$, $l = 1,\ldots,m$, consider its number of transitions

(i,j), $n_{\,i,j}(\mathbf{x}_l)$, such that $\sum_{i,j\in I} n_{\,i,j}(\mathbf{x}_l) = n_l - 1$, and the quantities $n_i^+(\mathbf{x}_l) = \sum_{j\in I} n_{\,i,j}(\mathbf{x}_l)$. We will consider the observed values $\{y_{\,0,0}(k)\}_{k=1,\dots,n_0^+(\mathbf{x}_l)}$ and $\{y_{\,1,1}(r)\}_{r=1,\dots,n_1^+(\mathbf{x}_l)}$ as realizations of the r.v.s $\{Y_{0,0}(k)\}_k$ and $\{Y_{1,1}(r)\}_r$. Note however that, even in the case when all the observed sequences $\{\mathbf{x}_1,\dots,\mathbf{x}_m\}$ have the same length, (i.e. $n_l = n$, $\forall\ l$), the quantities $n_i^+(\mathbf{x}_l)$ for each fixed $i \in \{0,1\}$, will not be, in general, constant amongst the various sequences. For that reason, reproducibility is a crucial property of the interaction model chosen, hence, we will not consider the multiplicative interactions, which are not reproducible. Note that the parameterizations in terms of the $\{p_{\,x_1,N}\}$ or $\{w_{\,x_1,N}\}$ presented in Section 2 are clearly not reproducible, while the parameterizations presented in Section 3 are reproducible, but it does not make sense to set some of their elements equal to zero. The mixture models presented in Section 2.1 are reproducible.

In case of ME variables, the additive interactions depend only on the numbers of variables involved for each exchangeable group $\{Y_{0,0}(k)\}_k$ and $\{Y_{1,1}(r)\}_r$, that is, we can define

$$Cov_{\,k,r} = E\left[\prod_{i=1}^{k}\Big(Y_{0,0}(i) - E[Y_{0,0}(i)]\Big)\prod_{j=1}^{r}\Big(Y_{1,1}(j) - E[Y_{1,1}(j)]\Big)\right]$$

By expanding the product, one can find the following formulas defining the one–to–one relation between parameters $\{m_{\,k,r}\}$ and $\{Cov_{\,k,r}\}$ (see [7]):

$$Cov_{\,k,r} = \sum_{i=0}^{k}\sum_{j=0}^{r}\binom{k}{i}\binom{r}{j}(-1)^{i+j}\left(m_{\,1,0}\right)^{i}\left(m_{\,0,1}\right)^{j}\ m_{\,k-i,r-j} \tag{10}$$

$$m_{\,k,r} = \sum_{i=0}^{k}\sum_{j=0}^{r}\binom{k}{i}\binom{r}{j}\left(m_{\,1,0}\right)^{i}\left(m_{\,0,1}\right)^{j}\ \ Cov_{\,k-i,r-j} \tag{11}$$

As a consequence, the parameterization in terms of the $\{Cov_{\,k,r}\}$, defined for every (k,r) such that $2 \le k+r \le n-1$, together with $m_{\,0,1}$, $m_{\,1,0}$ and q_1 represents all the n–ME distributions satisfying Ind.Ass. 1. ($Cov_{\,0,0}$ is always equal to 1 while $Cov_{\,1,0}$ and $Cov_{\,0,1}$ are always 0).

Note that the simplest case when all the $Cov_{\,k,r}$ are zero, represents the Markov chains models. In fact by (11), we would have $m_{\,k,r} = (m_{\,1,0})^k(m_{\,0,1})^r$, that is, by (9), a Markov chain having probability of transition $(0,0)$ equal to $m_{\,1,0}$, and probability of transition $(1,1)$ equal to $m_{\,0,1}$.

Recall that, in case of a mixture of Markov chains (2), the $\{m_{\,k,r}\}$ are the mixed moments of the mixing measure ν. Formulas (10) and (11) link the ordinary mixed moments and the central mixed moments of a bivariate distribution (see e.g. equations (34.28) (34.29) in [19]). Consequently, in a mixture of Markov chains model we have:

$$Cov_{\,k,r} = E_\nu\left[\left(\theta_{\,0,0} - E_\nu[\theta_{\,0,0}]\right)^k\ \left(\theta_{\,1,1} - E_\nu[\theta_{\,1,1}]\right)^r\right]$$

Then we can represent the mixtures of Markov chains models under Ind.Ass. 1 (for example the MBM) as particular additive models, specifically those where the $Cov_{k,r}$ are the mixed central moments of a measure over $[0,1]^2$. In particular, $Cov_{2,0}$ and $Cov_{0,2}$ respectively are the variance of $\theta_{0,0}$ and of $\theta_{1,1}$, and hence are necessarily nonnegative. That does not need to hold in general:

Example 1. *The 3–ME distribution defined by*

$$\begin{array}{llll} w_{0,\left(\begin{smallmatrix}2&0\\0&0\end{smallmatrix}\right)} = \frac{3}{32}, & w_{0,\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)} = \frac{2}{32}, & w_{0,\left(\begin{smallmatrix}1&1\\0&0\end{smallmatrix}\right)} = \frac{5}{32}, & w_{0,\left(\begin{smallmatrix}0&1\\0&1\end{smallmatrix}\right)} = \frac{6}{32} \\ w_{1,\left(\begin{smallmatrix}0&0\\0&2\end{smallmatrix}\right)} = \frac{2}{32}, & w_{1,\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)} = \frac{2}{32}, & w_{1,\left(\begin{smallmatrix}0&0\\1&1\end{smallmatrix}\right)} = \frac{6}{32}, & w_{1,\left(\begin{smallmatrix}1&0\\1&0\end{smallmatrix}\right)} = \frac{6}{32} \end{array}$$

satisfies Ind.Ass. 1, and by (1), (4), (8) *and* (10) *leads to*

$$m_{0,1} = m_{1,0} = \frac{1}{2}, \quad Cov_{2,0} = -\frac{1}{16} \quad Cov_{0,2} = Cov_{1,1} = -\frac{1}{8}$$

Actually, in a mixture model all the even order terms $Cov_{2k,2r}$ are necessarily positive. These restrictions on the parameters constitute a simple test for the extendibility of a ME sequence under Ind.Ass. 1, and can help us in deciding for an additive model instead of a mixture model. We can write many similar necessary conditions for extendibility using moments inequalities. One that has been found to be useful is the following: In a mixture model we necessarily have

$$m_{k,0} = E_\nu[\theta_{0,0}^{k}] \geq \left(E_\nu[\theta_{0,0}]\right)^k = m_{1,0}^{k} \qquad m_{0,r} = E_\nu[\theta_{1,1}^{r}] \geq \left(E_\nu[\theta_{1,1}]\right)^r = m_{0,1}^{r}$$

for any k and r. In particular, we can be interested in the values of $q_0\ m_{n-1,0}$ and $q_1\ m_{0,n-1}$ which are the probabilities of never changing the initial state. In a mixture those values can only exceed the corresponding values we would have in a simple Markov chain. On the converse, by (11) one can see that in an additive model, if the interaction parameters are negative, the inverse inequality can hold. In Example 1 we have

$$m_{2,0} = \frac{3}{16} < \frac{1}{4} = m_{1,0}^{2} \qquad \text{and} \qquad m_{0,2} = \frac{1}{8} < \frac{1}{4} = m_{0,1}^{2}$$

4.2. Estimation procedures

4.2.1. Saturated models

Say we have observed a sample of m binary sequences of length n, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, and let $\sharp[x_1, N]$ be the number of sequences in the sample starting with x_1 and consistent with the transition count matrix N, such that

$$\sum_{x_1 \in I} \sum_{N \in \Phi(x_1,n)} \sharp[x_1, N] = m$$

Then the likelihood and the log–likelihood of the sample under the parameterization $\{p_{x_1,N}\}$ respectively are:

$$L\Big(\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}\,;\,\{p_{\,x_1,N}\}\Big)=\prod_{x_1\in I}\ \prod_{N\in\Phi(x_1,n)}(p_{\,x_1,N})^{\sharp[x_1,N]} \tag{12}$$

$$l\Big(\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}\,;\,\{p_{\,x_1,N}\}\Big)=\sum_{x_1\in I}\ \sum_{N\in\Phi(x_1,n)}\sharp[x_1,N]\cdot\lg(p_{\,x_1,N}) \tag{13}$$

Obviously, in the same conditions the ML estimates $\{\widehat{w}_{\,x_1,N}\}$ of the $\{w_{\,x_1,N}\}$ simply are the observed proportions of sequences consistent with (x_1,N):

$$\widehat{w}_{\,x_1,N}=\frac{\sharp[x_1,N]}{m} \tag{14}$$

Since ML estimates are invariant under one–to–one functions, we can find the ML estimates of any saturated parameterization of an n–ME distribution by applying to the (14) the corresponding transformation. So, by (1) we can find the ML estimates of the $\{p_{\,x_1,N}\}$. Subsequently, by (4) we can obtain the ML estimates of the $\{w_{\,i,k,r}\}$. Then, under Ind.Ass. 1, by (8) we can obtain the ML estimates of the $\{m_{\,k,r}\}$, and by (10) the ML estimates of the $\{Cov_{\,k,r}\}$.

4.2.2. Reduced additive models

To construct the reduced additive models under Ind.Ass. 1 and to calculate the ML estimate of the parameters, first consider the log–likelihood (13) in terms of the $\{p_{\,x_1,N}\}$. Then, by formulas (5), (6), (7) and (11), we can rewrite it in terms of the $\{Cov_{\,k,r}\}$ together with $m_{\,0,1}$, $m_{\,1,0}$ and q_1. Thus we can construct a reduced additive model, simply by setting as zero in the log–likelihood expression some of the $Cov_{\,k,r}$, for example starting from those of higher orders, and finally find the ML estimates of the remaining parameters by numerically maximizing the log–likelihood with respect to them (note that the log–likelihood is always linear in the interaction parameters).

We can also construct the additive interactions without Ind.Ass. 1 defining

$$\frac{m_{\,i,k,r}}{q_i}=E_{P_i}\big[Y_{0,0}(1)\cdots Y_{0,0}(k)\cdot Y_{1,1}(1)\cdots Y_{1,1}(r)\big]\qquad i\in\{0,1\}$$

and

$$Cov_{\,i,k,r}=E_{P_i}\left[\prod_{h=1}^{k}\Big(Y_{0,0}(h)-E_{P_i}[Y_{0,0}(h)]\Big)\prod_{j=1}^{r}\Big(Y_{1,1}(j)-E_{P_i}[Y_{1,1}(j)]\Big)\right]$$

where P_i is the distribution of $(X_1,\ldots,X_n)$ conditioned on $\{X_1=i\}$ and E_{P_i} is the respective expectation. Then formulas (10) and (11) hold with $Cov_{\,i,k,r}$ replacing $Cov_{\,k,r}$ and $\frac{m_{\,i,k,r}}{q_i}$ replacing $m_{\,k,r}$. But, as we have said, without Ind.Ass. 1 we have an identifiability problem regarding the $\{m_{\,i,k,r}\}$. Then, any one–to–one transform of the $\{m_{\,i,k,r}\}$ would have the same problem, and we cannot define a saturated parameterization in terms of the $\{Cov_{\,i,k,r}\}$. Nevertheless, we can consider nearly all the corresponding reduced additive models.

In fact, if we a priori set as zero (at least) the parameters $Cov_{0,k,r}$ and $Cov_{1,k,r}$ for all the couples (k, r) such that $k + r = n - 1$, we can write the likelihood in terms of the $\{Cov_{0,k,r}\}$ and $\{Cov_{1,k,r}\}$ for $k+r < n-1$ together with the four parameters $\frac{m_{i,0,1}}{q_i}$, $\frac{m_{i,1,0}}{q_i}$, $i \in \{0, 1\}$ and numerically maximize it.

In case of a mixture model, we have some restrictions on the parameters analogous to those we have seen before under Ind.Ass. 1. We can consider some of them. In fact, even if the parameters $m_{i,k,r}$ are not identifiable without Ind.Ass. 1, the particular cases $m_{0,k,0}$ and $m_{1,0,r}$ (and hence $Cov_{0,k,0}$ and $Cov_{1,0,r}$) actually are. In fact, they correspond respectively to $p_{0,\left(\begin{smallmatrix}k&0\\0&0\end{smallmatrix}\right)}$ and $p_{1,\left(\begin{smallmatrix}0&0\\0&r\end{smallmatrix}\right)}$, so we can obtain them by (3). In a mixture we necessarily have

$$\frac{m_{0,k,0}}{q_0} \geq \left(\frac{m_{0,1,0}}{q_0}\right)^k \quad \text{and} \quad \frac{m_{1,0,r}}{q_1} \geq \left(\frac{m_{1,0,1}}{q_1}\right)^r \quad \forall\, k, r \tag{15}$$

Then, given a real dataset, starting from the (14), we can calculate the sample values $\widehat{m}_{0,k,0}$ and $\widehat{m}_{1,0,r}$, and eventually derive some evidences against a mixture model if the inequalities (15) do not hold.

5. Numerical examples

We now present an application of the models analyzed to three datasets. The first one has been obtained from a longitudinal study: the National Longitudinal Survey of Youth (NLSY79) whose data are freely available on web. Over 12000 persons answered a questionnaire by interview for several years. We have considered a variable concerning the labor status of the respondent during the week preceding the interview. The second one is a simulated dataset generated from a mixture of Markov chains model. The third, analyzed in [27], concerns the results of two kinds of medical tests on diabetic pregnant patients to determine fetal oxygenation.

On all the datasets, we have fitted a simple Markov chain (MC), a MBM model, some finite mixtures of Markov chains and some additive models (AM) with and without Ind.Ass. 1. To estimate the additive models, we have used a software that allows manipulation of mathematical expressions in symbolic form to write their log–likelihoods following the procedure described in Section 4.2.2. Then we utilized a Newton–Raphson maximization routine to maximize them. The ML estimates of the finite mixture models have been computed using the EM algorithm (for the equations utilized see [17] and [28]), while a numerical maximization routine has been used for the MBM model (see [27]). Since the fitted models have different numbers of parameters, we consider the Akaike and the Bayesian information criteria in order to compare their performances:

$$AIC = -2\,\hat{l} + 2c \qquad\qquad BIC = -2\,\hat{l} + c\ln(m)$$

where $\hat{l}$ is the value of the log–likelihood of the model corresponding to the ML estimates of the parameters, m is the number of sequences in the dataset, and c is the number of free parameters to be estimated in the model. Concerning the finite mixture models and the additive models, we present just the models that resulted in the best performance w.r.t. the two criteria. In particular, we have not checked all the possible additive models, so the results we show may be suboptimal.

5.1. NLSY dataset

We have dichotomized the variable concerning the labor status of the respondent (1=working, 0=not working) and considered a period of 12 years. We successively have excluded the units having missing values, finally obtaining a sample consisting of 5657 binary sequences of length 12. We consider them as independent realizations of a 12–ME sequence $(X_1, \dots, X_{12})$.

With regard to the finite mixture models, the fit is just slightly improved by adding more than 6 component Markov chains, and the models with 5 and 6 components have the best BIC and AIC respectively in their class.

The total number of possible additive models is huge. So we have restricted our attention to the lower order interaction parameters: $\{Cov_{\,i,k,r} : k+r \leq 6\}$, as it seems that in general they contribute more to the fit. In addition, to make a direct comparison with the mixture models, we have fixed a number of parameters equal to that of the best performing mixture models ($c = 19$ and 23). AM 1 in Table 1 was found to be the best additive model with those restrictions under Ind.Ass. 1. The others (AM 2 and AM 3) without Ind.Ass. 1.

AM 1 is defined by all the interaction parameters $Cov_{\,k,r}$, for k, $r \leq 3$ excluding $Cov_{\,2,2}$, together with $Cov_{\,4,0}$, $Cov_{\,0,4}$, $Cov_{\,4,2}$ and $Cov_{\,2,4}$. AM 2 is defined by the interaction parameters $Cov_{\,i,2,0}$, $Cov_{\,i,0,2}$, $Cov_{\,i,2,1}$, $Cov_{\,i,1,2}$ and $Cov_{\,i,4,0}$, for both $i = 0$ and $i = 1$, together with $Cov_{\,0,1,1}$, $Cov_{\,0,3,1}$, $Cov_{\,0,1,3}$ and $Cov_{\,0,0,4}$. AM 3 has all the parameters of AM 2 together with $Cov_{\,1,3,0}$, $Cov_{\,1,0,3}$, $Cov_{\,1,1,1}$ and $Cov_{\,0,2,3}$.

AM 1, AM 2 and AM 3 all have a better AIC and BIC than the mixture models. More generally, most of the additive models we have checked perform better than mixtures. To investigate such a clear difference we checked the

Table 1
Results for the NLSY data

	AM 1 (Ass 1)	**AM 2**	**AM 3**	**Saturated**
$-\hat{l}$	38635.9	38540.3	**38523.9**	38327.2
c	19	19	23	133
AIC	77309.8	77118.5	77093.9	**76920.4**
BIC	77435.9	**77244.7**	77246.7	77803.6

	M.C.	**MBM**	**Mix of 5 M.C.**	**Mix of 6 M.C.**
$-\hat{l}$	39146.9	38943.2	38657.1	38646.5
c	3	5	19	23
AIC	78299.8	77896.4	77352.3	77339.1
BIC	78319.8	77929.6	77478.5	77491.8

criteria for extendibility for the empirical distribution. Let $\widehat{m}_{0,k,0}$ be the sample value of $m_{0,k,0}$ for this dataset. We have that $\widehat{m}_{0,k,0}/\widehat{q}_0 - (\widehat{m}_{0,1,0}/\widehat{q}_0)^k$ is slightly positive for small values of k, it decreases with k and is negative for $k \geq 8$. We know that in the mixture models (15) holds, and in addition, all the fitted mixture models show an opposite trend to that of the empirical distribution, as the difference increases with k. Consequently, it is not surprising that the additive models in general work better than the mixtures for this dataset.

From the results obtained with the additive models we can derive some information on the data. Let us call the terms $Cov_{k,r}$ having both k and r non null, the "cross" terms. The MBM satisfies Ind.Ass. 2 that amounts to:

$$Cov_{k,r} = Cov_{k,0}\, Cov_{0,r} \qquad \forall\, k, r$$

Since $Cov_{0,1} = Cov_{1,0} = 0$, all the cross terms having k or r equal to one are null under the MBM, and in general the other cross terms are small in value. But the cross terms seem to be important for the additive models in this dataset, and that suggests a certain dependence between the two exchangeable subprocesses forming the ME process exists, i.e. Ind.Ass. 2 is untenable. That explains the inadequacy of the MBM for this dataset.

5.2. Simulated dataset

In [33] it is demonstrated that, in the space of all the n–exchangeable binary distributions, the proportion of them (with respect to the Lebesgue measure of the space) which are mixtures of i.i.d. model quickly tends to zero with n. That fact, together with the results in [7] suggest that an analogous proportion should be valid in the ME case, i.e. if we pick uniformly at random a distribution in the simplex of the parameters $w_{x_{1,N}}$ representing all the n–ME binary distributions, probably it will not be a mixture of Markov chains model. For that reason, and to test the versatility of the additive models, we simulated a dataset from a mixture of Markov chains. We have chosen a large family of mixtures, namely, a 8–components finite mixture whose parameters' values were randomly sampled with uniform probability over the parameters' space. The generated dataset consists of 100 binary sequences of length 6.

Model AM 1 in Table 2 satisfies Ind.Ass. 1, models AM 2 does not. AM 1 is defined by the interaction parameters $Cov_{2,0}$ and $Cov_{0,2}$. AM 2 is defined by $Cov_{0,2,0}$, $Cov_{1,2,0}$ and $Cov_{1,0,2}$.

Table 2
Results for the simulated data

	M.C.	**MBM**	**(Ass 1)** **AM 1**	**Mix of** **2 M.C.**	**AM 2**	**Mix of** **3 M.C.**	**Saturated**
$-\hat{l}$	399.14	384.33	385.07	385.39	382.17	**380.99**	378.92
c	3	5	5	7	8	11	31
AIC	804.28	**778.67**	780.14	784.78	780.34	783.98	819.84
BIC	812.11	**791.71**	793.17	803.02	801.18	812.63	900.61

The MBM results as the best model for both the criteria. We note however that, even in this adversely simulated dataset, the additive models are competitive as they show a performance comparable to that of the MBM, and preferable to those of the finite mixtures. In general, it is not hard to find an additive model with this characteristics for this dataset. The reason lies in the possibility of choosing an intermediate number of parameters, while in the finite mixtures we directly pass from $c = 7$ to $c = 11$.

In the various additive models checked, the cross terms did not show particular importance. AM 1 and AM 2 do not retain any cross term. That may suggest a weak dependence between the two exchangeable subprocesses, and that is consistent with the good performance of the MBM.

5.3. Fetal Oxygenation dataset

So far we have analyzed datasets consisting of sequences of equal lengths, but it is important to note that it is possible to use the additive models and the mixture models even if the sequences at hand have different lengths, since those models are reproducible. Say the lengths of the observed sequences range from n_{min} to n_{max}, then we simply have to consider a log–likelihood of the kind

$$\sum_{n=n_{min}}^{n_{max}} \sum_{x_1 \in I} \sum_{N \in \Phi(x_1,n)} \sharp[x_1, N] \cdot \lg\left(p_{\,x_1,N}\right)$$

instead of (13). The main difference with the case when all the sequences have the same length, is that the processing time of the algorithms increases, especially for the EM algorithm. For that reason, we now analyze a small dataset consisting of only 30 sequences whose lengths range from 2 to 7. Two kind of medical tests were repeatedly performed in different occasions on 30 units and for each occasion we fix a 1 if they resulted as concordant and a 0 if they resulted as discordant.

Model AM 1 in Table 3 satisfies Ind.Ass. 1; AM 2 does not. AM 1 is defined by parameters $Cov_{\,2,0}$ and $Cov_{\,0,2}$. AM 2 retains no interaction parameters, i.e. is just defined by $\frac{m_{\,0,0,1}}{q_0}$, $\frac{m_{\,0,1,0}}{q_0}$, $\frac{m_{\,1,0,1}}{q_1}$, $\frac{m_{\,1,1,0}}{q_1}$ and q_1. It substantially provides two transition probabilities matrices depending on whether the initial state is 0 or 1. Again, the additive models presented are preferable to the mixtures w.r.t. the AIC and BIC. In particular, the finite mixtures require too many parameters for this small dataset.

Table 3
Results for the Fetal Oxygenation data

			(Ass 1)		**Mix of**
	M.C.	**MBM**	**AM 1**	**AM 2**	**2 M.C.**
$-\hat{l}$	55.49	54.62	53.79	52.90	**52.88**
c	3	5	5	5	7
AIC	116.97	119.23	117.59	**115.81**	119.75
BIC	**121.17**	126.24	124.59	122.81	129.56

6. Conclusions

The mixtures of Markov chains do not represent all the ME distributions, and many cases not covered by them may be of practical interest. The models presented in this paper should fill the gap. The finite mixture models enjoy a great popularity mainly due to their neat interpretation in terms of cluster analysis. On the converse, the additive models suffer from the same problem of the log-linear models (or any interaction model) in terms of parameters selection and interpretability. Nevertheless, the additive models for ME data seem to be an interesting tool, their main strength lying in two points: First they represent an extremely flexible hierarchical class. In a finite mixture of Markov chains, we must add four parameters for every additional component. In an additive model we can consider any number of parameters ranging from a very simple model (a Markov chain, or even an i.i.d. model) to the saturated model. Moreover, for every fixed number of parameters we have several possible additive models among which we can choose. So, if we check a sufficiently large number of models (possibly all), we would likely find one resulting as preferable to any mixture, both in terms of goodness of fit, and in parsimony of parameters. The second advantage concerns the processing time of the estimation procedure. In fact, the ML estimates of the additive models are immediately obtained since Newton–Raphson methods, or similar devices, have a quadratic speed of convergence. That contrasts with the slowness of the EM algorithm which is a well–known problem. For example in the NLSY dataset the EM for the finite mixtures required several hours to converge and the processing time dramatically increases with the number of components.

Acknowledgements

The author wishes to thank Professor Rodolfo Rosa and Professor Luca Tardella for their support and helpful comments.

References

[1] Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statist.* **27** 162–167. MR506839

[2] Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in item analysis and prediction* 158–168. Stanford Univ. Press, Stanford, Calif. MR0121893

[3] Crisma, L. (1982). Quantitative analysis of exchangeability in alternative processes. In *Exchangeability in probability and statistics (Rome, 1981)* 207–216. North-Holland, Amsterdam. MR675976

[4] Darroch, J. N. (1974). Multiplicative and additive interaction in contingency tables. *Biometrika* **61** 207–214. MR0403087

[5] Darroch, J. N. and Speed, T. P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **11** 724–738. MR707924

[6] de Finetti, B. (1959). La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista. C.I.M.E., Induzione e Statistica (No.1), 1–115.

[7] Di Cecco, D. On the extendibility of Partially and Markov exchangeable binary sequences. arXiv:0908.4158v2 [math.PR]

[8] Diaconis, P. and Freedman, D. (1980). de Finetti's theorem for Markov chains. *Ann. Probab.* **8** 115–130. MR556418

[9] Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *Ann. Probab.* **8** 745–764. MR577313

[10] Diaconis, P. and Rolles, S. W. W. (2006). Bayesian analysis for reversible Markov chains. *Ann. Statist.* **34** 1270–1292. MR2278358

[11] Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied longitudinal analysis. Wiley Series in Probability and Statistics.* Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR2063401

[12] Fortini, S., Ladelli, L., Petris, G. and Regazzini, E. (2002). On mixtures of distributions of Markov chains. *Stochastic Process. Appl.* **100** 147–165. MR1919611

[13] Freedman, D. (1962). Invariants under mixing which generalize de Finetti's theorem. *Ann. Math. Statist* **33** 916–923. MR0156369

[14] Freedman, D. (1962). Mixtures of Markov processes. *Ann. Math. Statist.* **33** 114–118. MR0137156

[15] Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *J. Amer. Statist. Assoc.* **79** 632–638. MR763581

[16] Girolami, M. and Kabán, A. (2005). Sequential activity profiling: latent Dirichlet allocation of Markov chains. *Data Min. Knowl. Discov.* **10** 175–196. MR2145975

[17] Heckerman, D., Meek, C., Smyth, P. and White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Min. Knowl. Discov.* **7** 399–424. MR2011154

[18] Ip, E. H., Wang, Y. J. and Yeh, Y.-n. (2004). Structural decompositions of multivariate distributions with applications in moment and cumulant. *J. Multivariate Anal.* **89** 119–134. MR2041212

[19] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete multivariate distributions. Wiley Series in Probability and Statistics: Applied Probability and Statistics.* John Wiley & Sons Inc., New York. A Wiley-Interscience Publication. MR1429617

[20] Kupper, L. and Haseman, J. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34** 69-76.

[21] Lancaster, H. O. (1969). *The chi-squared distribution.* John Wiley & Sons Inc., New York. MR0253452

[22] Lindsay, B. G. (1995). *Mixture models: theory, geometry, and applications. NSF–CBMS Regional Conference Series in Probability and Statistics, vol.5.* IMS, Hayard, CA.

[23] Lipsitz, S. R., Fitzmaurice, G. M., Sleeper, L. and Zhao, L. (1996). Estimating the joint distribution of repeated binary responses: Some small sample results. *Comput. Stat. Data Anal.* **23** 219–227.

[24] Muliere, P., Secchi, P. and Walker, S. (2000). Urn schemes and reinforced random walk. *Stochastic Process. Appl.* **88** 59–78. MR1761692

[25] Pemantle, R. (2007). A survey of random processes with reinforcement. *Probab. Surv.* **4** 1–79 (electronic). MR2282181

[26] Quintana, F. A. and Newton, M. A. (1998). Assessing the order of dependence for partially exchangeable binary data. *J. Amer. Statist. Assoc.* **93** 194–202. MR1614608

[27] Quintana, F. A. and Newton, M. A. (1999). Parametric partially exchangeable models for multiple binary sequences. *Braz. J. Probab. Stat.* **13** 55–76. MR1788267

[28] Quintana, F. A. and Silva, A. (2006). Testing for differences among discrete distributions: an application of model-based clustering. *Braz. J. Probab. Stat.* **20** 141–152. MR2359225

[29] Stefanescu, C. and Turnbull, B. W. (2003). Likelihood inference for exchangeable binary data with varying cluster sizes. *Biometrics* **59** 18–24. MR2012138

[30] Streitberg, B. (1990). Lancaster interactions revisited. *Ann. Statist.* **18** 1878–1885. MR1074442

[31] Töwe, J., Bock, J. and Kundt, G. (1985). Interactions in contingency table analysis. *Biometrical J.* **27** 17–24. MR792539

[32] Whittle, P. (1955). Some distribution and moment formulae for the Markov chain. *J. Roy. Statist. Soc. Ser. B.* **17** 235–242. MR0077041

[33] Wood, G. R. (1992). Binomial mixtures and finite exchangeability. *Ann. Probab.* **20** 1167–1173. MR1175255

[34] Zabell, S. L. (1995). Characterizing Markov exchangeable sequences. *J. Theoret. Probab.* **8** 175–178. MR1308676

[35] Zaman, A. (1984). Urn models for Markov exchangeability. *Ann. Probab.* **12** 223–229. MR723741

[36] Zaman, A. (1986). A finite form of de Finetti's theorem for stationary Markov exchangeability. *Ann. Probab.* **14** 1418–1427. MR866363

[37] Zentgraf, R. (1975). A note on Lancaster's definition of higher-order interactions. *Biometrika* **62** 375–378. MR0378265