Multivariate Hidden Markov Models for disease progression

Andrea Martino[1], Giuseppina Guatteri[1] and Anna Maria Paganoni[1]

[1]Department of Mathematics, Politecnico di Milano, Milan, Italy

Abstract

Disease progression models are a powerful tool for understanding the development of a disease, given some clinical measurements obtained from longitudinal events related to a sample of patients. These models are able to give some insights about the disease progression through the analysis of patients histories and can be also used to predict the future course of the disease for an individual. In particular, Hidden Markov Models are suitable for disease progression since they model the latent unobservable states of the disease. In this work we introduce a novel HMM where the outcome is multivariate and its components are not independent; to accomplish our aim, since we do not make any usual normality assumptions, we model the outcome using copulas. We first test the performance of our model in a simulation setting and show the validity of the method. Then, we study the course of Heart Failure, applying our model to an administrative dataset from Lombardia Region in Italy, showing how episodes of hospitalization can give information about the disease status of a patient.
Keywords: Copulas; Disease progression; Hidden Markov Models; Multivariate data.

1 Introduction

Hidden Markov Models (HMMs) are a popular method for modeling disease progression and estimating the rates of transition between the stages of a disease, widely used in many fields including bioinformatics [7], signal processing [5], finance [23]. In the literature, there are many examples of HMMs used to model the progression of a chronic disease (see, e.g., the study of bronchiolits obliterans syndrome [19], glaucoma and Alzheimer's disease [21], HIV disease [15]). HMMs consist of a Markov model in which the underlying states visited by the Markov process are unobservable (i.e. hidden) but the distribution that generates the output depends on the state. Basically, a HMM can be considered as a generalization of a mixture model where the hidden variables, which control the mixture components,

are not independent of each other but related through a Markov process [25]. We only consider models where the state space of the hidden variables is discrete but the observations can either be generated by a discrete or continuous distribution; for example, like in many medical applications, the states of the Markov process may represent different severities of a disease, while the observations are some measure of a clinical index. In particular, in this work, we want to extend the use of HMMs to multivariate observations. In the multivariate HMM framework, there are examples of HMMs where the outcome is modelled using mixtures of multivariate Gaussian distributions [3] or, alternatively, using independent nonnormal distributions and it is possible to find `R` packages able to do the task [9, 28]. In general, assuming a multivariate normal distribution is not a suitable approach to model the joint distribution of certain variables. Therefore, to find a more appropriate multivariate model, we use copulas.

Copulas have been introduced in [26] but they received particular attention in statistical modelling only in the last two decades. A copula is a function that "couples" a multivariate distribution function to its marginal distribution functions and contains all the information about the dependence structure between the components of a random vector [22]. As a result, we can model the components of a random vector using different marginals by incorporating a flexible modelling of the dependence structure. For this reason, copulas are very appealing in many fields when the multivariate dependence is of interest and the multivariate normality is questionable; for instance, they are used in finance [4], medicine [8], engineering [14]. For further details about copulas and their theory, see among others [10, 22].

In this paper, we develop a novel HMM where the outcome is modelled using a mixture of multivariate distributions. The dependence structure between the continuous components of the outcome is described using the copula. Since our motivating problem regards disease progression, we apply our model to a dataset regarding patients in the Lombardia region of Italy affected by Heart Failure (HF), a degenerative pathology that interests the heart muscle. Specifically, the dataset is extracted from an administrative database and provides information about the hospital admissions of patients. For a detailed description of the dataset, along with a study using multi-state models, see [18]. The reason of our model choice consists in the fact that the dataset contains for each patient a sequence of observation vectors that cannot be modelled as realizations of a multivariate gaussian distribution and, since each element of the vector brings information about a single patient for a specific hospitalization, considering the presence of a dependence structure is the most obvious choice. Moreover, for each patient, we have a sequence of vectors that evolves in time and the HMMs are able to capture their progression through different stages of the disease.

The paper is organized as follows: in Section 2 we present the model, giving some background information about the theory of HMMs and copulas; in Section 3 we assess the performance of our model in a simulation study and in Section 4 we apply it to the administrative dataset; finally, Section 5 contains some discussion and conclusions. All the analysis have been carried out using the statistical software R [24] and the codes are available upon request.

2 The model

A Hidden Markov Model [12] is a bivariate process $\{(Q_k, \mathbf{X}_k)\}_{k\geq 0}$ defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

- $\{Q_k\}_{k\geq 0}$ is a Markov chain with a discrete and finite state space $\{s_1, \dots, s_N\}$, with $N \geq 1$, transition matrix $A = \{a_{ij}\} = \mathbb{P}(Q_k = s_j | Q_{k-1} = s_i)$ and initial distribution $\boldsymbol{\nu}$, where $\nu_i = \mathbb{P}(Q_0 = s_i)$;
- for each k, $\mathbf{X}_k$ is a d-dimensional random vector. Given the state process $\{Q_k\}_{k\geq 0}$, $\mathbf{X}_k$ is a sequence of conditionally independent random vectors; the conditional distribution of $\mathbf{X}_k$ only depends on Q_k for each k and admits a probability density function $f_{\mathbf{X}_k|Q_k}$.

For any $i = 1, \dots, N$, we denote with $b_i(\,\cdot\,; \boldsymbol{\theta}_i) = f_{\mathbf{X}_k|Q_k=s_i}(\,\cdot\,; \boldsymbol{\theta}_i)$ the emission probability density function of the random vector $\mathbf{X}_k$ conditionally on the event $\{Q_k = s_i\}$. Such density depends on some parameters $\boldsymbol{\theta}_i$ that belongs to a set $\Theta \in \mathbb{R}^L$, $L \geq 1$, e.g., if $f_{\mathbf{X}_k|Q_k=s_i}(\,\cdot\,; \boldsymbol{\theta}_i)$ is the pdf of an univariate exponential distribution, $\Theta = \mathbb{R}^+$ contains the rate parameter of the distribution. Hence, we can completely define our HMM with the set of parameters $\lambda = (\boldsymbol{\nu}, A, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$.

Let us indicate with $\mathbf{x} = \{\mathbf{x}_k, 1 \leq k \leq K\}$ a single realization of length K of the stochastic process $(\mathbf{X}_k)_k$; then we can denote with $\mathcal{L}(\lambda|\mathbf{x})$ the likelihood function of the parameters of the model λ given the data $\mathbf{x}$. There are usually three fundamental problems (see among others [25] and [11]) associated with HMMs:

1. find $\mathcal{L}(\lambda|\mathbf{x})$ for some observation $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_K)$;
2. given some $\mathbf{x}$ and λ, find the best state sequence $Q = (Q_1, \dots, Q_K)$ that explains $\mathbf{x}$;
3. find $\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, \mathcal{L}(\lambda|\mathbf{x})$.

As usually done in the literature, to address these problems we use the forward-backward procedure, the Viterbi algorithm and the Baum-Welch algorithm, respectively. In particular, since there are no modifications of the Viterbi algorithm due to our model, we only focus on the first and third problem. Let us recall the forward variables $\alpha_k(j)$ defined as

$$\alpha_k(j) \;=\; f_{(\mathbf{X}_1,\dots,\mathbf{X}_k,Q_k)|\lambda}(\mathbf{x}_1, \dots, \mathbf{x}_k, s_j),$$

i.e. the probability density function of observing the partial sequence $\mathbf{X}_1, \dots, \mathbf{X}_k$ and ending in the state Q_k, given the model λ. We can solve for $\alpha_k(j)$ inductively, as follows:

(1) initialization: $\alpha_1(j) = \nu_j b_j(\mathbf{x}_1; \boldsymbol{\theta}_j)$, for $1 \leq j \leq N$;

(2) induction: $\alpha_{k+1}(j) = \left[\sum_{i=1}^N \alpha_k(i) a_{ij}\right] b_j(\mathbf{x}_{k+1}; \boldsymbol{\theta}_j)$, for any $1 \leq k \leq K-1$ and $1 \leq j \leq N$;

(3) termination: $\mathcal{L}(\lambda|\mathbf{x}) = \sum_{i=1}^N \alpha_K(i)$.

Since these quantities are made up of products of probabilities, as k increases they become progressively smaller and eventually they are rounded to zero. In order to solve this problem, it is essential to introduce an appropriate scaling procedure, see for instance [25]. Hence, at each step k of the algorithm, we introduce a scaling parameter $c_k = 1/\sum_{i=1}^{N} \alpha_k(i)$, such that the scaled variable can be written as $\hat{\alpha}_k(j) = c_k \alpha_k(j)$. The backward procedure is similar; the backward variables $\beta_k(i)$ can be defined as

$$\beta_k(i) \;=\; f_{(\mathbf{X}_{k+1},\ldots,\mathbf{X}_K)|(Q_k,\lambda)}(\mathbf{x}_{k+1},\ldots,\mathbf{x}_K, s_i),$$

i.e. the probability density function of the partial observation sequence $\mathbf{X}_{k+1},\ldots,\mathbf{X}_K$ given that we started at state $Q_k = s_i$ and the model λ. We can solve for $\beta_k(i)$ inductively, as follows:

(1) initialization: $\beta_K(i) = 1$, for $1 \leq i \leq N$;

(2) induction: $\beta_k(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{x}_{k+1}; \boldsymbol{\theta}_j)\beta_{k+1}(j)$, for any $1 \leq k \leq K-1$, $1 \leq i \leq N$;

(3) termination: $\mathcal{L}(\lambda|\mathbf{x}) = \sum_{i=1}^{N} \beta_1(i)\nu_i b_i(\mathbf{x}_1; \boldsymbol{\theta}_i)$.

Again, it is essential to scale appropriately these variables. At each step k of the algorithm, we compute $\hat{\beta}_k(j) = c_k \beta_k(j)$, using the values of c_k computed in the forward step. Both procedures are able to compute $\mathcal{L}(\lambda|\mathbf{x})$ separately but we need them both in order to find the model λ that maximizes the likelihood.

Since the sequence of states occupied by the Markov-chain component of an HMM is unobservable, the usual approach consists in treating the states as missing data and apply an EM algorithm [6] to find the maximum likelihood estimates of the parameters. In the HMM framework, the algorithm is known as Baum-Welch algorithm (for further details see [1, 2, 29] for the case of discrete observations and [3] for the case of continuous observations).

In order to describe the procedure for the estimation of the HMM parameters, we first define $\xi_k(i,j)$, the probability of being in state s_i at time k, and state s_j at time $k+1$, given the model and the observation sequence, i.e.

$$\xi_k(i,j) \;=\; \mathbb{P}\left(Q_k = s_i, Q_{k+1} = s_j \mid X_1 = x_1, \ldots, X_K = x_K, \lambda\right)$$

We also define the probability of being in the state s_i at time k, given the observation sequence and the model

$$\gamma_k(i) \;=\; \mathbb{P}\left(Q_k = s_i \mid X_1 = x_1, \ldots, X_K = x_K, \lambda\right) \;=\; \sum_{j=1}^{N} \xi_k(i,j),$$

From [11], the log-likelihood of our model is

$$\begin{aligned}\log(\mathcal{L}(\lambda|\mathbf{x})) = &\underbrace{\sum_{j=1}^{N}\gamma_1(j)\log\nu_j}_{\text{term 1}} + \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\sum_{k=2}^{K}\xi_k(i,j)\right)\log a_{ij}}_{\text{term 2}} \\ &+ \underbrace{\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_k(j)\log f_{\mathbf{X}_k|Q_k=s_j}(\mathbf{x}_k;\boldsymbol{\theta}_j)}_{\text{term 3}}.\end{aligned} \tag{2.1}$$

Using this expression, it is possible to perform the EM algorithm for HMMs by following iteratively the two steps:

- **E step** replace the quantities $\xi_k(i,j)$ and $\gamma_j(k)$ by their conditional expectations given the current parameter estimates and the observations

$$\begin{aligned}\xi_k(i,j) &= \frac{\alpha_k(i)a_{ij}b_j(\mathbf{x}_{k+1};\boldsymbol{\theta}_j)\beta_{k+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_k(i)a_{ij}b_j(\mathbf{x}_{k+1};\boldsymbol{\theta}_j)\beta_{k+1}(j)}, \\ \gamma_k(i) &= \frac{\alpha_k(i)\,\beta_k(i)}{\sum_{j=1}^{N}\alpha_k(j)\,\beta_k(j)};\end{aligned}$$

- **M step** maximize the log-likelihood in (2.1). Since each term of the expression depends on different parameters, we can split it into three parts and maximize each term separately. The solutions which maximize each term are the following:

  1. $\bar{\nu}_i = \gamma_1(i)$, i.e., the expected number of times in state s_i at time $k = 1$;

  2. $\bar{a}_{ij} = \frac{\sum_{k=1}^{K-1}\xi_k(i,j)}{\sum_{k=1}^{K-1}\gamma_k(i)}$, i.e., the ratio between the expected number of transitions from state s_i to state s_j and the expected number of transitions from state s_i;

  3. the maximization of the third term depends on the assumptions for the state-dependent distributions. More details will be given at the end of the next section.

All the formulas presented until now only consider a single observation sequence. In order to have sufficient data to obtain reliable estimates of all model parameters, we need to use multiple sequences. Let us denote the set of observation sequences as

$$\mathcal{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$$

where M is the number of statistical units, each one of them being a sequence of length K_m, $m = 1, \ldots, M$, i.e., $\mathbf{x}^{(m)} = (\mathbf{x}_1^{(m)}\,\mathbf{x}_2^{(m)}\cdots\mathbf{x}_{K_m}^{(m)})$ is the m-th observation sequence of

length K. We assume all the sequences to be independent from each other; our goal is to adjust the parameters of the model λ to maximize the following likelihood:

$$\mathcal{L}(\lambda|\mathcal{X}) = \prod_{m=1}^{M} \mathcal{L}(\lambda|\mathbf{x}^{(m)}). \tag{2.2}$$

For more details about all the reestimation formulas of the HMM parameters, see [25].

2.1 The Copula Approach

We now focus on the study of the estimate of the emission probability density functions $b_1, \ldots, b_N$. Since we want to model the dependence among the components of each $\mathbf{X}_k$, if any, we construct our multivariate distribution using copula models. Differently from what is already available in the literature, in this work we want to consider a more general case, where the observations vectors are not necessarily gaussian and the components have a dependence structure. To fulfill this goal we create a probability model based on copulas, where the marginals are coupled into a joint distribution. First, we have to model the marginals by making some assumptions on the distribution family and then use a copula to take into account the dependence among the components (see [22]).

A d-dimensional copula is a function $C : [0,1]^d \to [0,1]$ with the following properties:

1. $\forall \mathbf{u} \in [0,1]^d, C(\mathbf{u}) = 0$ if any $u_i = 0$, for $i = 1, 2, \ldots, d$;
2. if all elements of $\mathbf{u}$ are 1 except u_i, then $C(\mathbf{u}) = u_i$;
3. $\forall \mathbf{u}, \mathbf{v} \in [0,1]^d$ such that $u_i < v_i$ for all i,

$$V_C([\mathbf{u}, \mathbf{v}]) \geq 0,$$

where $[\mathbf{u}, \mathbf{v}] = [u_1, v_1] \times [u_2, v_2] \times \ldots \times [u_d, v_d]$ and $V_C([\mathbf{u}, \mathbf{v}])$ is the nth order difference of C on $[\mathbf{u}, \mathbf{v}]$, i.e., $V_C([\mathbf{u}, \mathbf{v}]) = \Delta_{\mathbf{u}}^{\mathbf{v}} C(\mathbf{t}) = \Delta_{u_n}^{v_n} \Delta_{u_{n-1}}^{v_{n-1}} \cdots \Delta_{u_1}^{v_1} C(\mathbf{t}), \mathbf{t} \in [0,1]^d$, where the kth first order difference of the function C is defined as $\Delta_{u_k}^{v_k} C(\mathbf{t}) = C(t_1, \ldots, t_{k-1}, v_k, t_{k+1}, \ldots, t_n) - C(t_1, \ldots, t_{k-1}, u_k, t_{k+1}, \ldots, t_n)$.

For instance, if we take $d = 2$, the function $C : [0,1] \times [0,1] \to [0,1]$ is a copula if, for every $u, v \in [0,1]$, $C(u,0) = C(0,v) = 0$, $C(u,1) = u$, $C(1,v) = 1$ and for every $\mathbf{u}, \mathbf{v} \in [0,1] \times [0,1]$ such that $u_1 \leq u_2, v_1 \leq v_2$, for each hyperrectangle $[\mathbf{u}, \mathbf{v}]$ we have $V_C([\mathbf{u}, \mathbf{v}]) = C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$. See [22] for further details.

Let us recall Sklar's Theorem [22], which is one of the main results about copulas and represents the theoretical foundation for their application.

Theorem 2.1 (Sklar)**.** *Let F be a multivariate d-dimensional distribution function with the related marginal distribution functions $F_1, \ldots, F_d$. Then there exists a d-dimensional copula C such that for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$:*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{2.3}$$

If all the marginal distribution functions $F_1, \dots, F_d$ are continuous, then the copula function C is unique. Conversely, if C is a d-dimensional copula and $F_1, \dots, F_d$ are cumulative distribution functions, then the function F defined in (2.3) is a d-dimensional distribution function with marginals $F_1, \dots, F_d$.

Given d uniform marginal distributions $U_1 = F_1(X_1), \dots, U_d = F_d(X_d)$ of d random variables $X_1, \dots, X_d$ defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, if the marginal inverse distributions F_i^{-1} exist for all $i = 1, \dots, d$, we can write:

$$\begin{aligned} F(x_1, \dots, x_d) &= \mathbb{P}[X_1 \leq x_1, \dots, X_d \leq x_d] \\ &= \mathbb{P}[F_1^{-1}(U_1) \leq x_1, \dots, F_d^{-1}(U_d) \leq x_d] \\ &= \mathbb{P}[U_1 \leq F_1(x_1), \dots, U_d \leq F_d(x_d)] \\ &= C(F_1(x_1), \dots, F_d(x_d)). \end{aligned}$$

In this work we consider the case where each F_i is continuous and differentiable. Therefore, we can compute the copula density as

$$c = \frac{\partial^d C}{\partial F_1 \cdots \partial F_d}.$$

As a consequence of Sklar's Theorem, the joint density $f(x_1, \dots, x_d)$ can be written as:

$$f(x_1, \dots, x_d) = f_1(x_1) \cdots f_d(x_d) \cdot c[F_1(x_1), \dots, F_d(x_d)], \tag{2.4}$$

where $f_i(x_i)$ is the density corresponding to $F_i(x_i)$. This result states that the joint density, under appropriate conditions, can be written as a product of the marginal densities and the copula density.

In the literature, there are several classes of copulas, including the archimedean and the elliptical copulas. Archimedean copulas are very popular thanks to their tractability (see [13, 22]) while elliptical copulas are multivariate functions derived from elliptically countered distributions. Two common elliptical copulas are the Gaussian and Student's t. In this paper we focus on the d-dimensional Gaussian copula, since it is associated to a correlation matrix which allows us to model dependence and, differently from the archimedean copula, their marginal distributions are available analytically.

Let us denote with Γ the correlation matrix of the vector $\mathbf{x} \in \mathbb{R}^d$. Like other copula families, the gaussian copula allows any marginal distribution but similarly to the multivariate normal distribution it only consider pairwise dependence between individual components of a random vector. Let Φ denote the standard univariate normal distribution function while Φ_Γ represents the standard multivariate normal distribution function with correlation matrix Γ. Then the Gaussian copula can be written as

$$C_\Gamma(u_1, \dots, u_d) = \Phi_\Gamma(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$$

with the corresponding density [27]

$$c_\Gamma(u_1, \dots, u_d) = \det(\Gamma)^{-1/2} \exp\{-\frac{1}{2}\mathbf{q}^T(\Gamma^{-1} - I)\mathbf{q}\}. \tag{2.5}$$

where $\mathbf{q} = (q_1, \dots, q_d)^T$ and for all $i = 1, \dots, d$, $q_i = \Phi^{-1}(u_i)$, with $u_i = F_i(x_i)$. Using this result and (2.4), we can construct the desired joint density given any marginal density functions $f_1, \dots, f_d$ as:

$$f(x_1, \dots, x_d) = f_1(x_1) \cdots f_d(x_d) \cdot c_\Gamma(u_1, \dots, u_d) \tag{2.6}$$

Let us recall (2.1) and, in particular, the third term of the equation. As we stated in the previous section, its maximization depends on the assumptions made for the state dependent distributions. In the model we are considering each state distribution to emit an observation that can be modelled through copulas as a generic multivariate distribution. Since we are considering multiple observation sequences and we want to maximize the likelihood in (2.2), using the result obtained in (2.6), term 3 can be rewritten as:

$$\begin{aligned} \text{term 3} &= \sum_{m=1}^{M} \sum_{k=1}^{K_m} \sum_{j=1}^{N} \gamma_k(j) \log f_{\mathbf{X}_k | Q_k = s_k}(\mathbf{x}_k; \boldsymbol{\theta}_j) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K_m} \sum_{j=1}^{N} \gamma_k(j) \left(\log c_\Gamma[F_1(x_{1k}; \boldsymbol{\theta}_{1j}), \dots, F_d(x_{dk}; \boldsymbol{\theta}_{dj})] + \sum_{i=1}^{d} \log f_{ij}(x_{ik}; \boldsymbol{\theta}_{ij}) \right) \end{aligned} \tag{2.7}$$

where $f_{ij}(\,\cdot\,; \boldsymbol{\theta}_{ij})$ and $F_{ij}(\,\cdot\,; \boldsymbol{\theta}_{ij})$ are the univariate marginal pdf and cdf for component i and state j, respectively. Both the parameters of the marginal distributions $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_{1j}, \dots, \boldsymbol{\theta}_{dj})$, for all $j = 1, \dots, N$, and the correlation parameters contained in the matrix Γ belong to a set $\Theta \subset \mathbb{R}^L$, $L \geq 1$. To perform all the estimates and reduce the computational difficulty of the algorithm, we implement the two-stage estimation method called Inference Functions for Margins (IFM) proposed in [10]. Let us denote with $\boldsymbol{\rho}$ the vector containing all the correlation parameters of the correlation matrix Γ for the copula. In the first step we estimate the parameters $(\boldsymbol{\theta}_{lj})_{l=1,\dots,d;j=1,\dots,N}$ of the marginal distributions by computing for every state j and every component l

$$\widehat{\boldsymbol{\theta}}_{lj} = \underset{\boldsymbol{\theta}_{lj}}{\operatorname{argmax}} \sum_{m=1}^{M} \sum_{k=1}^{K_m} \gamma_k(j) \log f_l(x_{lk}; \boldsymbol{\theta}_{lj}).$$

To perform this step, we use the results presented in [17] and [20] and extend the estimators commonly used in the i.i.d. framework into the theory of the HMM. Specifically, let us denote with $\boldsymbol{\eta}$ a vector of parameters modelling the distribution of a sequence $\mathbf{X}_1, \dots, \mathbf{X}_K$ of i.i.d. random vectors and consider an estimator of $\boldsymbol{\eta}$ defined as $\widehat{\boldsymbol{\eta}} = \mathbf{g}(\sum_{k=1}^{K} \mathbf{X}_k, K)$ for some multivariate function $\mathbf{g}$. Then, we can extend these estimators to the HMM framework and write the estimator $\widehat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}_j$ based on the function $\mathbf{g}$ defined as follows:

$$\widehat{\boldsymbol{\theta}}_j = \mathbf{g} \left(\sum_{m=1}^{M} \sum_{k=1}^{K_m} \gamma_k(j)\, \mathbf{X}_k, \sum_{m=1}^{M} \sum_{k=1}^{K_m} \gamma_k(j) \right). \tag{2.8}$$

Given these estimates, in the second step, we can compute the correlation parameters $\boldsymbol{\rho}_j$, for all $j = 1, \ldots, N$, by

$$\widehat{\boldsymbol{\rho}_j} = \underset{\boldsymbol{\rho}_j}{\operatorname{argmax}} \sum_{m=1}^{M} \sum_{k=1}^{K_m} \gamma_k(j) \log c[F_1(x_{1k}, \widehat{\boldsymbol{\theta}}_{1j}), \ldots, F_d(x_{dk}, \widehat{\boldsymbol{\theta}}_{dj}); \boldsymbol{\rho}_j],$$

where we use the estimator proposed in (2.8).

3 Simulation Studies

In this section we present the results of a simulation study designed to investigate the performance of the copula model presented in the previous section (model A) and we compare it with a simpler HMM, where the components of the outcome are considered independent (model B). Moreover, we also consider some model selection criteria regarding the number of states for the HMM.

We generate a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $n = 5000$, in order to have $M = 250$ sequences of length $K = 20$. Each element of the sample is a realization of the joint distribution (X_1, X_2, X_3) where:

- $X_1 | s_j \sim Be(p_j)$;
- $\begin{pmatrix} X_2 \\ X_3 \end{pmatrix} = X_1 \mathbf{Y} + (1 - X_1)\mathbf{Z} = X_1 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + (1 - X_1) \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$, with $Y_{ij} = Y_i | s_j \sim \mathcal{E}(\lambda_{ij})$ and $Z_{ij} = Z_i | s_j \sim \mathcal{E}(\mu_{ij})$.

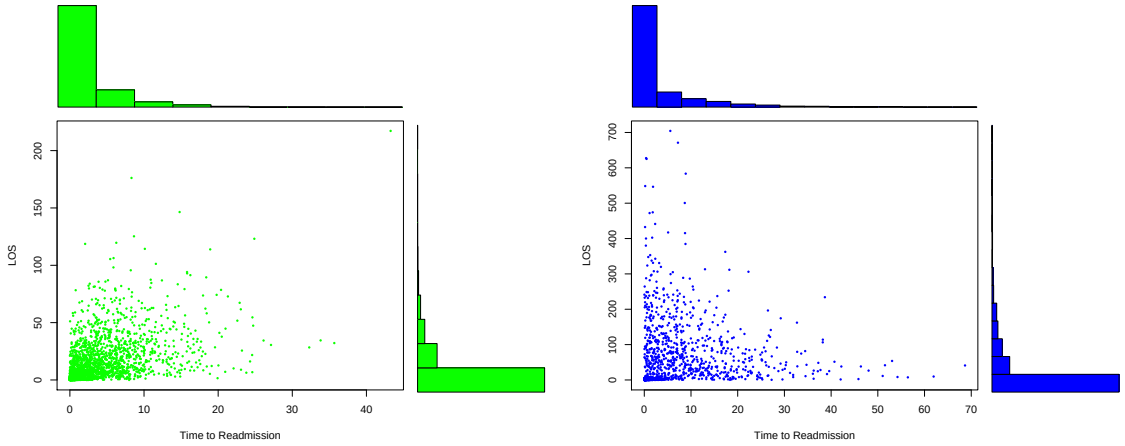
Note that, conditionally to the state, $X_1 \perp\!\!\!\perp (\mathbf{Y}, \mathbf{Z})$. For each statistical unit, the observations are the realization of a Markov process with three states, where State 3 is an absorbing state. Furthermore, given the state s_j, when generating the data we consider the components of the vectors $\mathbf{Y}$ and $\mathbf{Z}$ to be correlated with a correlation parameter $\rho_j = \operatorname{corr}(Y_{1j}, Y_{2j})$ and $\sigma_j = \operatorname{corr}(Z_{1j}, Z_{2j})$, respectively.

We use the following parameters to generate the data:

- **State 1**: $p_1 = 0.2, \lambda_{11} = 2, \lambda_{21} = 0.75, \mu_{11} = 3, \mu_{21} = 1, \rho_1 = 0.8, \sigma_1 = 0$.
- **State 2**: $p_2 = 0.7, \lambda_{12} = 1, \lambda_{22} = 0.5, \mu_{12} = 1.5, \mu_{22} = 0.75, \rho_2 = 0.2, \sigma_2 = -0.8$.
- **State 3**: $p_3 = 0.3, \lambda_{13} = 0.1, \lambda_{23} = 0.01, \mu_{13} = 0.2, \mu_{23} = 0.05, \rho_3 = -0.5, \sigma_3 = 0.5$.

$$\nu = (1 \quad 0 \quad 0), \qquad A = \begin{pmatrix} 0.70 & 0.20 & 0.10 \\ 0.15 & 0.80 & 0.05 \\ 0 & 0 & 1 \end{pmatrix}$$

In Fig. 1a and Fig. 1b we can see two plots along with the marginal histograms of the simulated dataset. Without using our prior knowledge, we assume the data to be marginally



(a) Simulated data for $x_1 = 0$ (b) Simulated data for $x_1 = 1$

a mixture of exponential distributions. Given the state s_j, the emission probability density function of an observation $\mathbf{x}_k = (x_1, x_2, x_3)_k$ for model A is computed as

$$\begin{aligned} b_j^A(\mathbf{x}_k; \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \rho_j, \sigma_j) \;=\; & \{p_j \cdot f_{1j}(y_{1k}; \lambda_{1j}) \cdot f_{2j}(y_{2k}; \lambda_{2j}) \cdot c[F_{1j}(y_{1k}; \lambda_{1j}), F_{2j}(y_{2k}; \lambda_{2j}); \, \rho_j]\}^{x_{1k}} \\ & \cdot \{(1 - p_j) \cdot f_{1j}(z_{1k}; \mu_{1j}) \cdot f_{2j}(z_{2k}; \mu_{2j}) \cdot c[F_{1j}(z_{1k}; \mu_{1j}), F_{2j}(z_{2k}; \mu_{2j}); \, \sigma_j]\}^{(1-x_{1k})} \end{aligned}$$

where f_{kj} and F_{kj}, $k = 1, 2$, are the pdf and the cdf of an exponential distribution, respectively. Starting from (2.7), taking into account that now we deal with a mixture of distributions, the third term of the likelihood to be maximized becomes:

$$\begin{aligned} (\text{term } 3)_A = & \sum_{m=1}^{M} \sum_{k=1}^{K_m} x_{1k} \cdot \sum_{j=1}^{N} \gamma_k(j) \Big(\log c[F_{1j}(y_{1k}; \boldsymbol{\lambda}_{1j}), F_{2j}(y_{2k}; \boldsymbol{\lambda}_{2j}); \, \rho_j] \\ & + \log p_j + \log f_{1j}(y_{1k}; \boldsymbol{\lambda}_{1j}) + \log f_{2j}(y_{2k}; \boldsymbol{\lambda}_{2j}) \Big) \\ & + \sum_{m=1}^{M} \sum_{k=1}^{K_m} (1 - x_{1k}) \cdot \sum_{j=1}^{N} \gamma_k(j) \Big(\log c[F_{1j}(z_{1k}; \boldsymbol{\mu}_{1j}), F_{2j}(z_{2k}; \boldsymbol{\mu}_{2j}); \, \sigma_j] \\ & + \log (1 - p_j) + \log f_{1j}(z_{1k}; \boldsymbol{\mu}_{1j}) + \log f_{2j}(z_{2k}; \boldsymbol{\mu}_{2j}) \Big) \end{aligned}$$

For model B, the emission probability density function is instead computed as

$$\begin{aligned} b_j^B(\mathbf{x}_k; \boldsymbol{\lambda}_j, \boldsymbol{\mu}_j, \rho_j, \sigma_j) \;=\; & \{p_j \cdot f_{1j}(y_{1k}; \lambda_{1j}) \cdot f_{2j}(y_{2k}; \lambda_{2j})\}^{x_{1k}} \\ & \cdot \{(1 - p_j) \cdot f_{1j}(z_{1k}; \mu_{1j}) \cdot f_{2j}(z_{2k}; \mu_{2j})\}^{(1-x_{1k})} \end{aligned}$$

where the joint density function is simply the product of the marginals, because of the

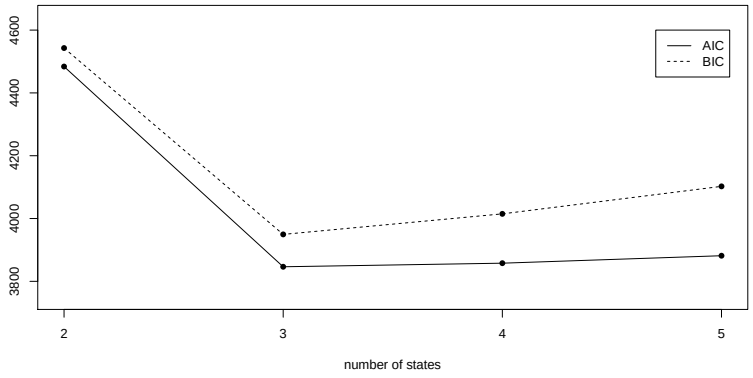


Figure 2: AIC and BIC for our model applied to the simulated data with $N = 2, \ldots, 5$ states.

independence between the two variables. The relative term to maximize for model B is

$$\begin{aligned}(\text{term } 3)_B = &\sum_{m=1}^{M}\sum_{k=1}^{K_m} x_{1k} \cdot \sum_{j=1}^{N} \gamma_k(j) \left(\log p_j + \log f_{1j}(y_{1k}; \boldsymbol{\lambda}_{1j}) + \log f_{2j}(y_{2k}; \boldsymbol{\lambda}_{2j})\right) \\ &+ \sum_{m=1}^{M}\sum_{k=1}^{K_m} (1 - x_{1k}) \cdot \sum_{j=1}^{N} \gamma_k(j) \left(\log (1 - p_j) + \log f_{1j}(z_{1k}; \boldsymbol{\mu}_{1j}) + \log f_{2j}(z_{2k}; \boldsymbol{\mu}_{2j})\right)\end{aligned}$$

Before investigating the results obtained for the parameter estimates, we run the Baum-Welch algorithm using the model presented in the previous section with $N = 2, \ldots, 5$ states to select the appropriate number of states, i.e., the "order" of the HMM. To this end, we compute the AIC and BIC for each HMM [11]. In Fig. 2 we can see a plot of the results of the two model criteria plotted against the number of states. Both AIC and BIC exhibit the lowest value for $N = 3$ states which represent the 'optimal' number of states.

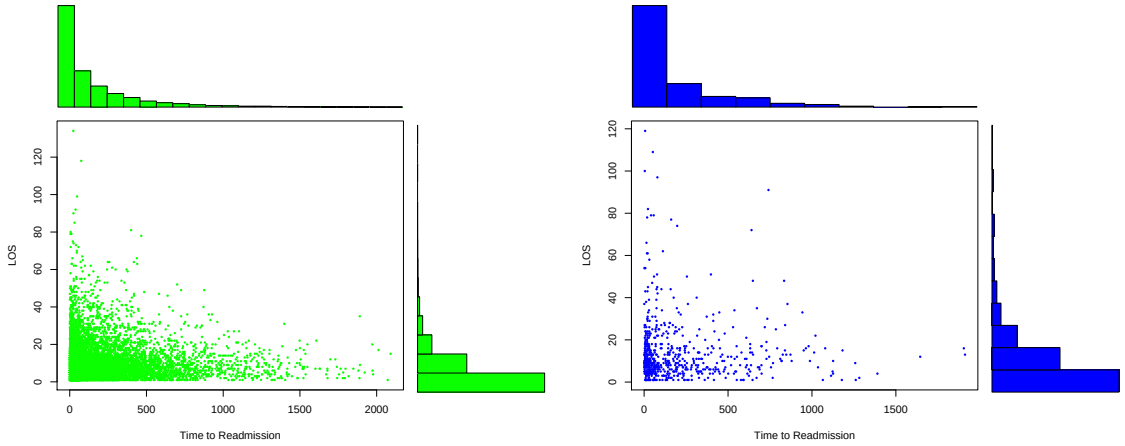
We run 50 replications of the Baum-Welch algorithm on the simulated data and then compute the Average Mean Square Error and standard deviation of the results, which are reported in Tab. 1. Both models are able to identify the presence of an absorbing state, for this reason we don't show the result for the estimates of the third row of matrix A, since the estimates are accurate for both models. If we compare all the results, it is clear that model A is able to give much more precise estimates of all the parameters used to generate the data, both in terms of mean value and standard deviation. Moreover, with model A we can also obtain accurate estimates of the correlation parameters used to generate the data.

4 Case study: disease progression

The data we use in this study belongs to a Heart Failure project and was extracted from the administrative warehouse of Regione Lombardia. All the observations are records

Table 1: A.M.S.E. (S.D.) of the HMM parameters for 50 simulation runs of the Baum-Welch algorithm with $N = 3$ states for model A and model B.

Parameter	Model A	Model B
a_{11}	$4.08 \cdot 10^{-4}$(0.0213)	$9.04 \cdot 10^{-3}$ (0.0558)
a_{12}	$3.04 \cdot 10^{-4}$ (0.0203)	$8.24 \cdot 10^{-3}$ (0.0382)
a_{13}	$8.85 \cdot 10^{-5}$ (0.0129)	$1.34 \cdot 10^{-4}$ (0.0093)
a_{21}	$1.33 \cdot 10^{-3}$ (0.0249)	$2.95 \cdot 10^{-2}$ (0.0827)
a_{22}	$1.4 \cdot 10^{-3}$ (0.0266)	$2.77 \cdot 10^{-2}$ (0.0792)
a_{23}	$6.08 \cdot 10^{-5}$ (0.0111)	$8.62 \cdot 10^{-5}$ (0.0131)
p_1	$6.01 \cdot 10^{-4}$ (0.0155)	$6.01 \cdot 10^{-4}$ (0.0270)
p_2	$1.20 \cdot 10^{-3}$ (0.0185)	$7.7 \cdot 10^{-1}$ (0.0653)
p_3	$8.31 \cdot 10^{-5}$ (0.0080)	$3.02 \cdot 10^{-1}$ (0.0072)
λ_{11}	$6.90 \cdot 10^{-2}$ (0.2624)	12.01 (1.3959)
λ_{21}	$4.39 \cdot 10^{-3}$ (0.0956)	$6.87 \cdot 10^{-2}$ (0.4919)
μ_{11}	$1.8 \cdot 10^{-2}$ (0.1117)	$1.90 \cdot 10^{-2}$ (0.1346)
μ_{21}	$8.98 \cdot 10^{-4}$ (0.0335)	$1.87 \cdot 10^{-2}$ (0.0588)
λ_{12}	$1.49 \cdot 10^{-3}$ (0.0424)	1.90 (1.0137)
λ_{22}	$1.49 \cdot 10^{-3}$ (0.0186)	$1.03 \cdot 10^{-2}$ (0.0241)
μ_{12}	$3.10 \cdot 10^{-2}$ (0.0797)	$2.51 \cdot 10^{-1}$ (0.3171)
μ_{22}	$3.02 \cdot 10^{-3}$ (0.0562)	4.92 (1.4875)
λ_{13}	$2.37 \cdot 10^{-5}$ (0.0025)	$1.33 \cdot 10^{-5}$ (0.0035)
λ_{23}	$1.52 \cdot 10^{-7}$ (0.0003)	$3.34 \cdot 10^{-7}$ (0.0004)
μ_{13}	$1.92 \cdot 10^{-5}$ (0.0045)	$3.29 \cdot 10^{-5}$ (0.0041)
μ_{23}	$1.88 \cdot 10^{-6}$ (0.0012)	$1.42 \cdot 10^{-6}$ (0.0013)
ρ_1	$7.45 \cdot 10^{-5}$ (0.0647)	—
ρ_2	$4.76 \cdot 10^{-6}$ (0.034)	—
ρ_3	$5.77 \cdot 10^{-6}$ (0.0223)	—
σ_1	$2.56 \cdot 10^{-7}$ (0.0192)	—
σ_2	$8.41 \cdot 10^{-5}$ (0.0231)	—
σ_3	$6.72 \cdot 10^{-6}$ (0.0244)	—



(a) HF data for patients with IC=0

(b) HF data for patients with IC=1

of hospital admissions of a patient, collected in a data warehouse, called SDO (Scheda di Dimissione Ospedaliera, i.e., hospital discharge paper) database and contains all the HF episodes with subsequent hospitalizations. Moreover, it is possible to retrieve information about both the hospitalizations (diagnoses and procedures, date of admission and discharge, vital status at discharge, ...) and the patients (sex, date of birth,...). For the patients who died before the end of study, the date of death was obtained through database linkage with the Italian National Registry of deaths (for further details about the dataset, see [18]).

In order to describe the relation between hospital admissions, length of hospitalizations and mortality of the patients, we adopt a HMM to describe how an individual moves between a series of discrete states in time. For each patient and each hospitalization we consider the following three outputs for our model: the time to readmission from a previous hospitalization, the length of stay (LOS) in hospital and a binary variable indicating if the patient was admitted to intensive care (IC) during the stay at the hospital. We consider a sample of 2,248 patients, corresponding to 11,039 observations, containing hospitalizations from 2006 to 2012 and excluding all the events with less than four hospitalizations recorded, in order to be able to see at least one transition between two states, if any. Among these individuals, 1,284 (57.11%) died by the end of the study.

The number of admission to hospital per patient ranged between 1 and 28. The mean (standard deviation) LOS is 10.8 (10.1) days (min = 1, median = 8, first and third quartiles respectively equal to 4 and 14, max = 134 days) while the mean (standard deviation) Time to Readmission is 207.3 (271.8) days (min = 1, median = 102, first and third quartiles respectively equal to 32 and 274, max = 2,091 days). Moreover, among all the hospitalizations, only for 817 (7.4%) of them the patient was submitted to intensive care.

In Fig. 3a and Fig. 3b we plotted the data, with the Time to Readmission to hospital on the horizontal axis and the LOS on the vertical axis, on the left for patients who did not undergo to intensive care and on the right for the patients who did. We can consider both variables to be marginally mixtures of exponential distributions. Before fitting the data to

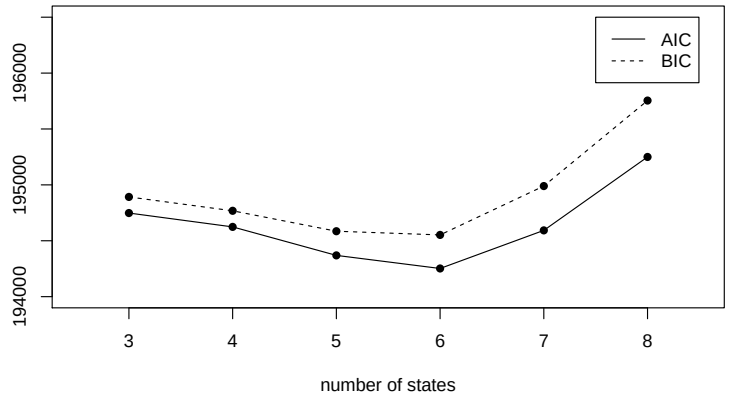


Figure 4: AIC and BIC for our model applied to the HF data with $N = 3, \dots, 8$ states.

our model, we perform the Hoeffding's independence test [16] to justify if there is global dependence between the two variables. Since we obtain a low p-value for the test, 0.0484, we can state that there is some dependence between the two variables and we can proceed with the application of our model.

Since the real number of states is not a priori known for our data, we apply the model described in Section 3, using as number of states $N = 3, \dots, 8$, and we compute AIC and BIC, which are plotted against the number of states in Fig. 4. Both criteria suggest that $N = 6$ represents the optimal number of states for this dataset, where we have by construction 5 transients states and 1 absorbing state which corresponds to death.

In Tab. 2 we show the results obtained for the chosen number of states. Specifically, we can detect two "groups" of states among the transient ones: group A, represented by states 1, 2 and 4, contains all the patients with a less advanced stage of the pathology, contrarily to the patients belonging to state 3 and 5, that belong to group B. In fact, the probability of starting from group B is null, as it is the probability of going to the death state from a state of group B. Moreover, we can notice how the probability p of being in intensive care is higher when belonging to a state of the group A. It is interesting to look at the values representing the average LOS. If we compare $\lambda_{2\cdot}^{-1}$ and $\mu_{2\cdot}^{-1}$ we can see how the estimated value of the LOS is in general higher for the patients admitted to intensive care but it is definitely higher for those belonging to group B. Finally, it is also worth noticing that the values of the correlation parameters are usually close to zero or negative meaning that, conditionally to the state, if the time to readmission increases the LOS tends to decrease, so patients who spend more time in hospital tend to be hospitalized more often.

Table 2: Results of the Baum-Welch algorithm applied to the HF data for the HMM with $N = 6$ states.

Parameter	State 1	State 2	State 3	State 4	State 5	State 6
$\nu_{\cdot}$	0.7420	0.0652	0.0000	0.1928	0.0000	0.0000
$p_{\cdot}$	0.0508	0.0085	0.1276	0.0806	0.1377	0.3509
$a_{1\,\cdot}$	0.5691	0.0321	0.2425	0.0600	0.0963	0.0000
$a_{2\,\cdot}$	0.0020	0.8308	0.0410	0.0013	0.1250	0.0000
$a_{3\,\cdot}$	0.1954	0.0465	0.0055	0.0081	0.0247	0.7197
$a_{4\,\cdot}$	0.4789	0.0112	0.1077	0.3017	0.1005	0.0000
$a_{5\,\cdot}$	0.0000	0.0354	0.0212	0.0521	0.3599	0.5314
$\lambda_{1\cdot}^{-1}$	309.5748	106.2818	301.9847	23.3091	42.1138	—
$\lambda_{2\cdot}^{-1}$	12.3933	13.1573	36.1757	11.5663	18.9560	—
$\rho_{\cdot}$	0.0222	-0.5847	0.0282	-0.1530	-0.0773	—
$\mu_{1\cdot}^{-1}$	280.2974	96.5228	291.9966	25.9869	35.0322	—
$\mu_{2\cdot}^{-1}$	9.8776	11.9321	10.1254	10.5595	12.4020	—
$\sigma_{\cdot}$	0.0088	0.1280	0.0467	-0.1010	-0.1403	—

5 Discussion and future developments

In this paper we propose a novel HMM to fit multivariate data, where the components of the outcome have a dependence structure, which is modelled by means of copulas. The model is very general, since it allows the choice of any marginal distributions and does not need to consider any normality assumptions on the data. In the simulation study we first showed how this model is a considerable improvement over the HMMs whose outcome has independent components and is able to estimate with high precision all the parameters of the model. Then, we applied the new model to an administrative longitudinal database concerning patients affected by HF. The obtained results provide an interesting insight at the latent disease progression.

In the context of HMMs, some generalizations can be done to enrich the model; for instance, in this work we only considered the case with discrete time, but it could be interesting to investigate a more general and complex case, where the HMM evolves in continuous time and can jump from a state to another one at any time. Another improvement can be done by modifying the model and adding some covariates, which would help to have a better understanding of some problems, both in the prior and the transition model. Finally, some more developments can be done when modeling the dependence structure by considering different kinds of copulas and then choosing the best one using some goodness-of-fit criteria. For example, by considering a Student's t or an Archimedean copula, it would be possible to take into account data with different dependency structures.

References

[1] Baum, L. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3, 1-8.

[2] Baum, L. E., Petrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1), 164-171.

[3] Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510), 126.

[4] Cherubini, U., Luciano, E., Vecchiato, W. (2004). Copula methods in finance. *John Wiley & Sons.*

[5] Crouse, M. S., Nowak, R. D., Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on signal processing*, 46(4), 886-902.

[6] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

[7] Durbin, R., Eddy, S. R., Krogh, A., Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. *Cambridge university press.*

[8] Eban, E., Rothschild, G., Mizrahi, A., Nelken, I., Elidan, G. (2013). Dynamic copula networks for modeling real-valued time series. *Artificial Intelligence and Statistics (pp. 247-255).*

[9] Jackson, C. H. (2011). Multi-state models for panel data: the msm package for R. *Journal of statistical software*, 38(8), 1-29.

[10] Joe, Harry. Joe, H. (1997). Multivariate models and multivariate dependence concepts. *Chapman and Hall/CRC.*

[11] Zucchini, W., MacDonald, I. L., Langrock, R. (2016). Hidden Markov models for time series: an introduction using R. *Chapman and Hall/CRC.*

[12] Cappé, O., Moulines, E., Rydén, T. (2009). Inference in hidden markov models. *Proceedings of EUSFLAT Conference (pp. 14-16).*

[13] Genest, C., Rivest, L. P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American statistical Association*, 88(423), 1034-1043.

[14] Genest, C., Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4), 347-368.

[15] Guihenneuc-Jouyaux, C., Richardson, S., Longini Jr, I. M. (2000). Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline. *Biometrics*, 56(3), 733-741.

[16] Hoeffding, W. (1948). A non-parametric test of independence. *The annals of mathematical statistics*, 546-557.

[17] Juang, B. H. (1985). Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T technical journal*, 64(6), 1235-1249.

[18] Ieva, F., Jackson, C. H., Sharples, L. D. (2017). Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology. *Statistical methods in medical research*, 26(3), 1350-1372.

[19] Jackson, C. H., Sharples, L. D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in medicine*, 21(1), 113-128.

[20] Liporace, L. (1982). Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, 28(5), 729-734.

[21] Liu, Y. Y., Li, S., Li, F., Song, L., Rehg, J. M. (2015). Efficient learning of continuous-time hidden markov models for disease progression. *Advances in neural information processing systems (pp. 3600-3608).*

[22] Nelsen, R. B. (2007). An introduction to copulas. *Springer Science & Business Media.*

[23] Paas, L. J., Vermunt, J. K., Bijmolt, T. H. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4), 955-974.

[24] R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

[25] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.

[26] Sklar, A. (1959). Functions de repartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris*, 8, 229-231.

[27] Song, P. X.-K.(2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2), 305-320.

[28] Visser, I., Speekenbrink, M. (2010). depmixS4: an R package for hidden Markov models. *Journal of Statistical Software*, 36(7), 1-21.

[29] Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4), 10-13.