Hidden Markov Models for disease progression
Hidden Markov Models per la progressione di patologia

Andrea Martino, Andrea Ghiglietti, Giuseppina Guatteri, Anna Maria Paganoni

Abstract Disease progression models are a powerful tool for understanding and predicting the development of a disease, given some longitudinal measurements obtained from a sample of patients. These models are able to give some insights about the disease progression through the analysis of patients histories and could be also used to predict the future course of the disease in an individual. In particular, Hidden Markov Models (HMMs) are a useful tool for disease modeling since they allow to model situations where the state of the disease is not observable, by giving the possibility to incorporate some priors and constraints. We applied our models to a simulated dataset by considering a generalization of HMMs with continuous time and multivariate outcome.

Abstract *I modelli di progressione di patologia sono un potente strumento per comprendere e prevedere lo sviluppo di una patologia, date delle misurazioni longitudinali ottenute da un campione di pazienti. Questi modelli sono in grado di arricchire la conoscenza sulla progressione della patologia attraverso l'analisi della storia dei pazienti e possono anche essere usati per predire il corso futuro della patologia di un individuo. In particolare, gli Hidden Markov Models (HMMs) sono un utile strumento per la modellazione di patologia in quanto permettono di costruire un modello in situazioni in cui lo stato della patologia non è osservabile, dando la possibilità di incorporare delle prior e dei vincoli. Abbiamo applicato i nostri modelli a un dataset simulato considerando una generalizzazione degli HMM a tempo continuo e risposta multivariata.*

Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni
Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133, Milan, Italy
e-mail: andrea.martino@polimi.it
e-mail: giuseppina.guatteri@polimi.it
e-mail: anna.paganoni@polimi.it

Andrea Ghiglietti
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123, Milan, Italy
e-mail: andrea.ghiglietti@unicatt.it

Key words: disease progression modeling, Hidden Markov Models, multivariate data

1 Introduction

Many chronic diseases can be naturally represented in terms of staged progression. Hidden Markov Models (HMMs) are a popular method for modeling disease progression and estimating the rates of transition between the stages of a disease. For this reason, we introduce a HMM for disease progression which takes into account the possibility of modeling multivariate observations with correlated components. Although discrete-time HMMs are often used to model disease progression, they are not very suitable in practice because the measurement data should be regularly sampled at discrete intervals and state transitions can only occur at these discrete times. Since we are interested in using our model to study the Heart Failure (HF) pathology and hospitalizations for HF patients occur irregularly in time, we consider a continuous time HMM, in which both the transitions between the hidden states and the observations can occur at arbitrary continuous times (for further details, see [1, 2]).

2 The model

A continuous time HMM is very suitable for modeling disease progression, which is a continuously evolving process. Even though the continuous time HMM adds more flexibility to the models with respect to the discrete time HMM, this comes with a higher computational cost. Indeed, in this case not only the hidden states are unobserved but the transition times are unknown too. Moreover, although HMMs often consider the case where the observations are sampled from a discrete distribution which can only take a finite number of values, in our analysis we have considered observations composed by both continuous and discrete values.

Let us denote $(\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_K)$ the set of observations, univariate or multivariate data, observed at K irregularly-distributed continuous points in time $(t_1, t_2, \ldots, t_K)$, such that we have two levels of hidden informations: the state of the Markov chain is hidden and the state transitions are also hidden, since it is not known if other transitions occur between two consecutive observations. For any observation $\mathbf{O}_k$ we denote the probability of being in a state $s(t_k)$ at time t_k, often called as *emission probability*, as $p(\mathbf{O}_k|s(t_k))$. As for continuous time Markov chains, we can define the finite and discrete state space S, the state transition rate matrix Q and the initial state probability distribution π. The elements q_{ij} in the matrix Q represent the rate of a process's transition from state i to state j for $i \neq j$, while the elements q_{ii} must be specified such that each row of the matrix sums up to zero ($q_i = \sum_{j \neq i} q_{ij}, q_{ii} = -q_i$) [1]. Moreover, if the process is time-homogeneous, the sojourn time in each state i

is exponentially distributed with parameter q_i, while q_{ij}/q_i indicates the probability of the next transition of the process from state i to state j.

As done in [3], if we consider a continuous time HMM which is fully observed, the complete joint likelihood of the data can be written as

$$CL = \prod_{k'=0}^{K'} q_{y_{k'},y_{k'+1}} e^{-q_{y_{k'}}\tau_{k'}} \prod_{k=0}^{K} p(\mathbf{O}_k|s(t_k)) = \prod_{i=1}^{|S|} \prod_{j=1,j\neq i}^{|S|} q_{ij}^{n_{ij}} e^{-q_i\tau_i} \prod_{k=0}^{K} p(\mathbf{O}_k|s(t_k))$$

where $(t'_0,t'_1,\ldots,t'_{K'})$ are the K' state transition times with $Y' = \{y_0 = s(t'_0),\ldots,y_{K'} = s(t'_{K'})\}$ being the corresponding states of the Markov chain, $\tau_k = t_{k+1} - t_k$ is the time interval between two observations while $\tau_{k'} = t'_{k'+1} - t'_{k'}$ is the time interval between two transitions and n_{ij} is the number of transitions from state i to state j.

3 Parameter estimation

For the computation of the rate transition matrix Q, several methods based on EM algorithms have been proposed in [3]. We now focus on the estimation of the emission probability matrix $B = \{b_j(\mathbf{O}_k)\}$ where $b_j(\mathbf{O}_k) = p(\mathbf{O}_k|s(t_k) = j)$ is the probability of being in a state j at time t_k, while observing $\mathbf{O}_k$. This estimation is usually straightforward if we are considering an univariate outcome. In general, since we consider multivariate observations, we also want to model the correlation among the variables.

As usually done, if the observations are only symbols chosen from a finite alphabet, a discrete probability density can be used to model the data and estimate the matrix B. This approach is not generally enough since we consider observations coming from both discrete and continuous distributions. In order to use a continuous observation density, we have to consider some restrictions for the estimation of the probability density function (pdf). We can represent the pdf in its most general representation as a finite mixture which can be written as

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm}(\mathbf{O}) \mathscr{D}[\mathbf{O}, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}], \quad 1 \leq j \leq N$$

where $\mathbf{O}$ is the vector being modeled, N is the number of statistical units, $c_{jm}(\mathbf{O})$ is a non negative mixture coefficient for the mth mixture in state j and $\mathscr{D}$ is a pdf with mean vector $\boldsymbol{\mu}_{jm}$ and covariance matrix $\mathbf{U}_{jm}$ for the mth mixture component in state j. We implemented a normalized version of the forward-backward algorithm and a Baum-Welch algorithm (see [5]) by using our matrix B in a continuous time framework. All the analysis have been carried out using the software `R` ([4]).

4 Simulation study

We applied our model to multivariate longitudinal data, which are repeated observations of multiple response variables. Since the data are correlated over time and multiple responses are measured at the same time, special treatments are required to analyze the data. In particular, the easiest approach would be to ignore the correlation, which would lead to some loss of information. Therefore, to flexibly characterize the distribution of the emission probabilities, we modelled the correlation among the observation components for each multivariate outcome.

We generated a sample $(x_1, y_1), \ldots, (x_N, y_N)$ of $N = 1000$ observations for $n = 50$ statistical units, in order to have 20 observations for each statistical unit. For each one of them, we have a sequence of pairs which is the realization of a 3-state Markov process. Given the state j of the Markov process, each pair of the sample is a realization of the joint distribution (X, Y), where $X \sim Be(p)$ and $Y = XY_1 + (1-X)Y_2$, with $Y_i \sim N(\mu_{ij}, \sigma_{ij}^2)$ independent of X. We used the following parameters to generate the data:

- **State 1**: $p = 0.2, \mu_{11} = 0, \mu_{21} = 3, \sigma_{11} = 0.5, \sigma_{21} = 0.8$.
- **State 2**: $p = 0.9, \mu_{12} = 1, \mu_{22} = 5, \sigma_{12} = 0.5, \sigma_{22} = 0.8$.
- **State 3**: $p = 0.7, \mu_{13} = 4, \mu_{23} = 7, \sigma_{13} = 0.5, \sigma_{23} = 0.8$.

We applied our algorithm using $m = 2, \ldots, 5$ states and we can see the results we obtained in Table 1. A problem which naturally arises is that of selecting an appropriate model, e.g. of choosing the appropriate number of states for the HMM, so we need some criteria for model comparison. To address this problem, for each run of the algorithm we computed the Aikake Information Criterion as $\text{AIC} = -2\log L + 2p$ and the Bayesian Information Criterion as $\text{BIC} = -2\log L + p\log T$, where L is the likelihood function of the fitted model, p is the number of unknown parameters and T is the number of observations. The values we obtained are showed in Fig. 1. According to both AIC and BIC, the model with three states is the most appropriate. As we can see for $m = 3$, the estimated values are very similar to the real ones, so we can conclude that by considering the correlation among the components of the outcome variables, we were able to obtain very good results.

5 Conclusion

In this work we built a model for longitudinal data with multivariate observations and showed that, if we consider the correlation among the components of the outcome, we can obtain very good results. The next step will consist in applying our algorithm to a real case study, with the data coming from an administrative dataset about hospitalizations which is in pre-processing, in order to study the progression of the Heart Failure pathology.

$m=2$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.4834	0.8298	3.1500	1.0470	1.2354
State 2	0.6634	3.9745	6.9547	0.4628	0.7692
$m=3$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.1981	0.0421	2.9863	0.5422	0.8521
State 2	0.9026	1.0127	4.9409	0.5335	0.7823
State 3	0.7067	3.9732	6.9780	0.4727	0.8324
$m=4$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.3003	0.0460	2.9873	0.4157	0.7113
State 2	0.1507	1.0128	4.8392	0.4067	0.6854
State 3	0.9120	3.7699	6.7472	0.5332	0.7222
State 4	0.6694	4.3680	7.2856	0.4727	0.7341
$m=5$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.1431	0.1124	2.6973	0.4131	0.7034
State 2	0.2937	0.7995	2.9427	0.2844	0.5854
State 3	0.3110	1.0190	4.3863	0.7944	0.9653
State 4	0.9143	3.8483	6.9001	0.5287	0.7712
State 5	0.6684	4.3742	7.2486	0.4728	0.7338

Table 1: Results of a HMM with $m = 2,\ldots,5$ states.

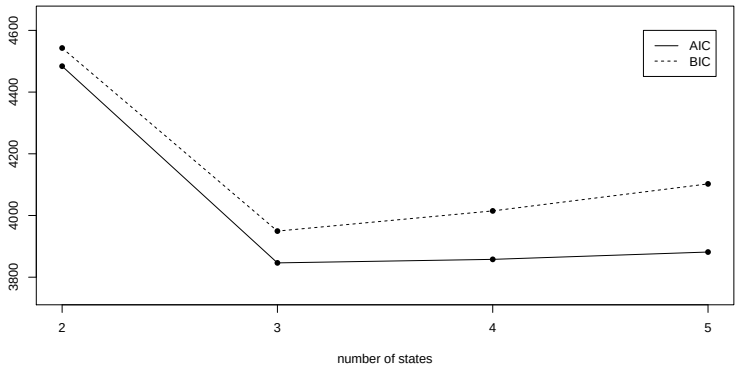

Fig. 1: Model selection criteria using AIC and BIC

References

1. Cox DR, Miller HD. The Theory of Stochastic Processes. Chapman and Hall; London: 1965.
2. Jackson CH. Multi-state models for panel data: the msm package for R. Journal of Statistical Software. 2011;38 (no. 8)
3. Y.Y. Liu, S. Li, F. Li, L. Song, J.M. Rehg, Efficient Learning of Continuous-Time Hidden Markov Models for Disease Progression, Advances in Neural Information Processing Systems, 3599-3607
4. 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/.
5. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE, **77**, 257–285 (1989)