Discrete or Continuous-Time Hidden Markov Models for Count Time Series (⋆)

Modelli Hidden Markov in tempo discreto o continuo per serie storiche di conteggio

Lorena CM Viviano
Dipartimento di Ingegneria dell'Informazione e Metodi Matematici
Università degli Studi di Bergamo
e-mail: lorena.viviano@unibg.it

Keywords: hidden markov models, embeddability

1. Introduction

In Hidden Markov Models (HMM) the probability distribution of response Y_t ($\forall t = 1, 2, \ldots, T$) at each observation time is conditionally specified on the current hidden or latent state X_t. The sequence of hidden states follows a first order time-homogeneous Markov chain. Discrete time or continuous time HMM are respectively specified by $T \subseteq \mathbb{N}$ or $T \subseteq \mathbb{R}^+$ (from now on DHMM and CHMM). In this work we compare some different goals of DHMM and CHMM. An application to bathing water quality data is considered.

2. Discrete or continuous

HMM are characterized by: the number of states of the hidden variable X_t, the initial state distribution $\pi_0 = P(X_1 = i)$ and the probability distributions of the observations $f(Y_t|X_t; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. What distinguishes DHMM form CHMM is the transition probability matrix $\mathbf{P}$ with elements $p_{ij} = P(X_t = j|X_{t-1} = i)$ ($p_{ij} \geq 0$ and $\sum_i p_{ij} = 1$) for DHMM and the transition intensity matrix $\mathbf{Q}$ with elements $q_{ij} = \lim_{\Delta t \to 0}[p_{ij}(t, t + \Delta t)]/[\Delta t] \quad \forall i \neq j$ ($q_{ii} = -\sum_{i \neq j} q_{ij}(t)$) for CHMM.

DHMM represent the classical approach referring to HMM (MacDonald and Zucchini, 1997). Considering, for example, the automatic speech recognition in engineering or the gene identification in DNA sequences in biology, a discrete step structure is the only logical approach. In such case time is not relevant or present and an equally spaced step is used conventionally for the underlying ordered sequence of states. CHMM concern to phenomena evolving in continuous time but that are usually observed at discrete irregularly spaced times. Examples come from panel studies in medical sciences, in which a patient state of health is monitored in time. Observations are made at irregular time points while the (hidden) disease progression evolves in continuous time. Some diseases are characterized by different states before recovery or death; the exact times of transition among states can not be known. Usually the main interest is to make inference on transition intensities in the different disease states or to compute the mean sojourn time in each state, rather than to estimate the transition probabilities. For this reason time plays a central role in CHMM and each sojourn time has an exponential distribution.

(⋆) Work supported by PRIN project 2006131039

The transitions probabilities are function of time and of transition intensities. The matrix $\mathbf{P}(t)$ evaluated in an interval time $(0, t)$ is obtained form the Kolmogorov differential equations as $\mathbf{P}(t) = \exp(t\mathbf{Q})$ (Taylor and Karlin, 1994); in some simple cases it is possible to compute the probabilities exactly. DHMM can be obtained by discrete time sampling of CHMM. On the other hand, $\mathbf{Q}$ can be achievable from $\mathbf{P}$ computed in DHMM only if embeddability conditions are satisfied (Kalbfleish and Lawless, 1985). Further complications arise if a vector of covariates $\boldsymbol{z}$ affects the transitional intensities, so that they are function of a vector of unknown parameters $\boldsymbol{\gamma}$: $q_{ij} = \exp(\boldsymbol{\gamma}_{ij}^T \boldsymbol{z})$. Theorems on embeddability are given in Israel *et al.* (2001).

In this work we focus on a count time series. The dataset analyzed comes from a study of bathing water quality in the district of Savona. Data were collected from April 2001 to September 2001. The response variable is the 'Number of Fecal Streptococcus' (counts in 100 ml of water), that presents an excess of zero counts ($\approx 64\%$). CHMM are applied since the state of water pollution evolves in continuous time and observations are done at discrete unequally spaced time points for economical, practical and sampling design reasons. A two states model is considered to model the excess of zeroes, with the hypothesis that $X_t = 0$ determines zero counts, while $X_t = 1$ determines the presence of fecal streptococcus (modeled by the Negative Binomial distribution). In this way, zero-inflation count data are modeled in a dynamic context.

The analysis was conducted in R with the library `msm` (Jackson, 2007). Among some other fitted models with more covariates, a simple model with two explanatory variables (oxygen and water temperature) affecting the distributions of counts gave a good fitting (AIC criterion) and produced the following estimated intensity matrix:

$$\hat{\mathbf{Q}} = \begin{bmatrix} -0.1107 & 0.1107 \\ 2.115 & -2.115 \end{bmatrix}$$

Since each sojourn time has an exponential distribution, the expected value of the sojourn time in the same state is $-1/q_{ii}$. The mean sojourn time in $X_t = 0$ is higher than the mean sojourn time in $X_t = 1$, meaning that when an area is characterized by the absence of pollution this condition is kept longer than what happens for the polluted condition.

References

Israel R.B., Rosenthal J.S., Wei J.Z. (2001) Finding generators for Markov chains via empirical transition matrices, with application to credit ratings, *Mathematical Finance*, 11, 245–265.

Jackson C. (2007) *Multi-state modelling with* R*: the* `msm` *package,* Cambridge, United Kingdom.

Kalbfleish J.D., Lawless J.F. (1985) The analysis of panel data under a Markov assumption, *Journal of the American Statistical Association*, 80, 863–871.

MacDonald I.L., Zucchini W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series,* Chapman & Hall, Great Britain.

Taylor H.M., Karlin S. (1994) *An Introduction to Stochastic Modeling,* Academic Press.