Hidden Markov models: Theory, applications and new perspectives

Fulvia Pennoni
Department of Statistics and Quantitative Methods
University of Milano-Bicocca
Email: fulvia.pennoni@unimib.it

Joint works with

Francesco Bartolucci

Silvia Pandolfi

University of Perugia

Luca Brusa
University of Milano-Bicocca

Outline

- Hidden Markov models
- Algorithm for variable and model selection
- Latent potential outcomes
- Tempered expectation maximization algorithm

Introduction

- Hidden Markov models are used to formulate complex dependence relations among observable variables, accounting for unobserved heterogeneity, and cluster units in separate groups
- These models nowadays have an important role in handling the complexity of modern data
- They find application in the analysis of both time-series and longitudinal categorical data in many different fields of interest especially, economics and medicine where complex data structures typically arise
- The main assumption underlying these models is that the observed data depend on a latent process that follows a Markov chain, typically of first-order, which may be homogeneous or heterogeneous over time

Hidden Markov model

- With reference to longitudinal categorical data let $\boldsymbol{Y}^{(t)} = (Y_1^{(t)}, \ldots, Y_r^{(t)})'$ denote the occasion-specific response variables for each time occasion $t = 1, \ldots, T$ and let $\boldsymbol{Y}$ denote the column vector of responses
- Each variable $Y_j^{(t)}$, $j = 1, \ldots, r$, $t = 1, \ldots, T$, is categorical with c categories
- Let $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})'$ denote the latent process having a discrete distribution with k states
- Model parameters are initial probabilities, denoted by $\pi_u = p(U^{(1)} = u)$, $u = 1, \ldots, k$, and transition probabilities denoted by $\pi_{u|\bar{u}}^{(t)} = p(U^{(t)} = u | U^{(t-1)} = \bar{u})$, $t = 2, \ldots, T$, $\bar{u}, u = 1, \ldots, k$

Hidden Markov model

- Model parameters include conditional response probabilities, denoted by $\phi_{jy|u}$, $u = 1, \ldots, k$, $j = 1, \ldots, r$, $y = 0, \ldots, c-1$

- The model in its basic formulation relies on the following three main assumptions:
  - $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$ are conditionally independent given $\boldsymbol{U}$;
  - $Y_1^{(t)}, \ldots, Y_r^{(t)}$ are conditionally independent given $U^{(t)}$, for $t = 1, \ldots, T$
  - $\boldsymbol{U}$ follows a first-order Markov chain with state space $1, \ldots, k$, where k is the number of latent states

- These assumptions are suitably relaxed according to the applicative context

Hidden Markov model

- Extended versions of the Hidden Markov model account for time-varying unobserved heterogeneity
- Observed response variables can be effectively summarized when these are multivariate
- Once the model is estimated it allows for accurate predictions at individual level
- Maximum likelihood approach to estimating the model parameters is based on the complete data log-likelihood function and it is performed through the Expectation-Maximization (EM) algorithm

Hidden Markov model

- Suitable recursions (forward and backward) are able to reduce the computational burden of estimation strongly
- Prediction of the sequence of latent states: local decoding is performed to predict the subject-specific sequence of latent states, which is based on the estimated posterior probabilities of U_{it} directly provided by the EM algorithm
- `LMest` is an available package implementing a general framework for a variety of hidden Markov models for categorical and continuous data with missing values and for data having a hierarchical structure in the `R` language

Algorithm for variable and model selection

- Hidden Markov models represent a useful tool for performing model-based clustering in order to group a set of individuals into distinct groups or clusters
- They allow for dynamic model-based clustering, and they may be seen as an extension of the latent class approach
- We propose a general method for model and variable selection for repeated continuous data when there are missing values under the missing at random assumption
- The implemented greedy forward-backward search algorithm is aimed at selecting a subset of relevant variables for clustering according to the Bayesian Information Criterion (BIC, Schwarz, 1978) and jointly the selection of the number of groups similar to the proposal of Raftery and Dean (2006) in the context of finite mixture models

- It starts with an initial number of states and of variables properly chosen and it repeatedly compares two models $\mathcal{M}_1$ and $\mathcal{M}_2$, each including or excluding a candidate variable

- At the h-th iteration, the EM algorithm performs an *Inclusion step*, and *Exclusion step* to achieve a trade-off between most smallest BIC for modeling and choosing the most informative predictors, and again *Model selection step*

- The BIC index under model $\mathcal{M}_1$ is expressed as

$$BIC(\mathcal{M}_1) = BIC_k(\mathcal{Y} \cup j) + BIC_{reg}(\bar{\mathcal{Y}} \setminus j \sim \mathcal{Y} \cup j)$$

  - j is the candidate variable, $\mathcal{Y}$ be the set of initial selected clustering variables, $\bar{\mathcal{Y}}$ the set of remaining variables
  - BIC_k is computed under the proposed hidden Markov model
  - BIC_{reg} is computed under a multivariate linear regression of the remaining variables, $(\bar{\mathcal{Y}} \setminus j)$ on the set of $\mathcal{Y} \cup j$ serving as approximation

- BIC index for model M2, in which j is not used for clustering, is expressed as

$$BIC(\mathcal{M}_2) = BIC_k(\mathcal{Y}) + BIC_{reg}(j \sim \mathcal{Y}) + BIC_{reg}(\bar{\mathcal{Y}} \setminus j \sim \mathcal{Y} \cup j)$$

- Difference between BIC of models $\mathcal{M}_1$ and $\mathcal{M}_2$ (in which j is not used for clustering) is considered to decide including or excluding a variable

$$BIC_{diff} = BIC(\mathcal{M}_1) - BIC(\mathcal{M}_2)$$

- *Inclusion step*: each variable j in the remaining set of variables $\bar{\mathcal{Y}}^{(h-1)}$, is singly proposed for inclusion in $\mathcal{Y}^{(h)}$
  - ◇ The variable with the smallest negative BIC_{diff} is included in $\mathcal{Y}^{(h-1)}$

Inclusion-exclusion algorithm

- *Exclusion step*: each variable j in $\mathcal{Y}^{(h)}$ is singly proposed for the exclusion
  - ◇ The variable with the highest positive value of the BIC_{diff} is removed from $\mathcal{Y}^{(h)}$
- *Model selection*: the current value of $k^{(h-1)}$ is updated by minimizing the BIC_k index of the HM model for the current set of clustering variables $\mathcal{Y}^{(h)}$ over k, from $(k^{(h-1)} - 1)$ to $(k^{(h-1)} + 1)$, in order to obtain the new value of $k^{(h)}$
- The algorithm ends when no more variables are added to or removed from $\mathcal{Y}^{(h)}$

Latent potential outcomes

Causal hidden Markov model

- Hidden Markov models may account for certain forms of unobserved confounding and thus can be used for causal inference on the treatment of interest
- We proposed a new formulation of the model based on potential versions of the latent variables related to the idea of potential outcomes as proposed in Rubin (1974) and later on extended by Holland (1986)
- Rosenbaum (1987), Imbens (2000), Robins et al. (2000), and Robins (2003) introduced the inverse-probability-of-treatment weighted (IPTW) estimator aimed to remove selection bias through statistical models

Causal hidden Markov model

- Notation:
  - Let $\boldsymbol{Y}_{it}$ be a column vector of r binary response variables defined for every individual i collected at each time occasion $t, t = 1, \ldots, T$
  - Let Z_i, $i = 1, \ldots, n$, be a categorical variable indicating the treatment for each individual i, $i = 1, \ldots, n$, with levels from 0 to $l - 1$
- Latent potential outcomes are defined as individual-and time-specific latent variables $H_{it}^{(z)}$, with $i = 1, \ldots, n$, $t = 1, \ldots, T$, having a discrete distribution with support points
- According to the consistency rule, $H_{it} = H_{it}^{(z_i)}$, where z_i is the observed treatment of individual i

Causal hidden Markov model

- Assuming that the set of pre-treatment covariates $\boldsymbol{V}_i$ is sufficiently informative: Z_i is independent from $H_{it}^{(z)}$ given $\boldsymbol{V}_i$ for $i = 1, \ldots, n$
- The proposal differs from the standard PO approach since $H_{it}^{(z)}$ and H_{it} are never directly observable
- We assume the stable unit treatment value assumption and positivity (Angrist et al., 1996), that is, $0 < P(Z_i = 1|\boldsymbol{V}_i) < 1$ for $i = 1, \ldots, n$
- A multinomial logit model is considered for $\log \frac{p(Z_i=z|\boldsymbol{V}_i)}{p(Z_i=1|\boldsymbol{V}_i)}$ individuals with missing responses at some time points

- Given a column vector of the time-varying post-treatment covariates $\boldsymbol{X}_{it}$ and assuming exogeneity, the initial probabilities are parametrized as

$$\log \frac{p(H_{i1}^{(z)} = h|\boldsymbol{X}_{i1} = \boldsymbol{x})}{p(H_{i1}^{(z)} = 1|\boldsymbol{X}_{i1} = \boldsymbol{x})} = \alpha_h + \boldsymbol{d}(z)'\boldsymbol{\beta}_{1h} + \boldsymbol{x}'\boldsymbol{\beta}_{2h}, \quad h = 2, \ldots, k,$$

  - α_h is an intercept specific for each latent state
  - $\boldsymbol{\beta}_{1h} = (\beta_{1h2}, \ldots, \beta_{1hl})'$ is a column vector of $l - 1$ regression parameters referred to the treatment levels

- Since each element β_{1hz} of $\boldsymbol{\beta}_{1h}$ for $z > 0$, is a shift parameter from the first logit with respect to the logit h, each of these parameters can be interpreted as the Average Treatment Effect on the initial probabilities
- Dynamic average treatment effects are estimated by adopting a similar parameterization on the transition probabilities

Tempered EM algorithm

Tempered Expectation-Maximization algorithm

- To account for the problem of multimodality of the model likelihood function, we implemented two tempered versions of the EM algorithm
- The likelihood is typically multimodal, and this implies that there is uncertainty about whether the solution at convergence of the maximum likelihood estimation algorithm is the optimal one
- The typical solution to this problem consists in trying different starting values for the estimation algorithm on the basis of deterministic and stochastic random rules (Maruotti and Punzo, 2021)
- Tempering or simulated annealing techniques consist on rescaling the objective function depending on a parameter, known as temperature, which controls the prominence of global and local maxima (Sambridge 2014)

Tempered Expectation-Maximization algorithm

- Let $\ell^*(\boldsymbol{\theta})$ denote the complete data log-likelihood function, the EM algorithm alternate the following steps until a suitable convergence condition:
  - E-step: compute the conditional expected value of $\ell^*(\boldsymbol{\theta})$, given the observed data and the value of the parameters at the previous step
  - M-step: maximize the expected value of $\ell^*(\boldsymbol{\theta})$ and so update the model parameters
- We implement the tempered EM (T-EM) algorithm by adjusting the computation of the conditional expected frequencies in the E-step $q(\cdot)$ on the basis of a temperature that controls the prominence of local maxima

Tempered Expectation-Maximization algorithm

- By properly tuning the sequence of temperature values, the procedure is gradually attracted towards the global maximum, escaping local sub-optimal solutions:
  - high temperatures allow exploring wide regions of the parameter space, avoiding being trapped in non-global maxima
  - low temperatures guarantee a sharp optimization in a local region of the solution space
- We define a sequence of temperatures $(\tau_h)_{h \geq 1}$, such that:
  - τ_1 is sufficiently small so that $\tilde{q}^{(\tau_1)}(\cdot)$ is relatively flat
  - τ_h tends towards 1 as the algorithm iteration counter increases

Main References

- Bartolucci F, Farcomeni A, Pennoni F. (2013) *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman and Hall/CRC.
- Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, **81**, 1-38.
- Maruotti, A. and A. Punzo (2021). Initialization of hidden Markov and semi-hidden Markov: a critical evaluation of several strategies. *Int. Stat. Rev.*, **89**, 447–480.
- Raftery, A. E., and Dean, N. (2006). Variable selection for model-based clustering. J. Ame. Stat. Ass., **101**, 168–178.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical* , **100**, 322–331.
- Sambridge, M. (2014). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geo- phys. J. Int.*, **196**, 357–374.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**: 461-464.