A Hidden Markov Approach to the Analysis of Incomplete Multivariate Longitudinal Data

Francesco Lagona

1 Dipartimento di Scienze Politiche, Università degli Studi di Roma Tre, (e-mail: francesco.lagona@uniroma3.it)

Keywords: Hidden Markov model, Mini mental state examination, Missing values, multivariate longitudinal data.

1 Introduction

Multivariate longitudinal data provides a unique opportunity in studying the joint evolution of multiple response variables over time. These data are typically examined by random effects models, specified on the basis of continuous (often normally distributed) random effects, which account for the dependence between the outcomes.

A limitation of this approach is that latent variables, in the form of subject-specific parameters, are time-constant. Hidden Markov models (HMMs) represent a flexible alternative for modeling multiple outcomes which are repeatedly measured over time (Bartolucci et al., 2013; Bartolucci et al., 2009). The basic assumption of these models is that the response variables, which are longitudinally observed, are conditionally independent given the states visited by a latent Markov chain which allows for time-varying unobserved heterogeneity.

In the case of complete data information, maximum likelihood estimation (MLE) of HMMs can be obtained by appropriate E-M algorithms. When the data are incomplete, i.e. some of the outcomes are missing, the missing mechanism should be taken into account, in order to obtain efficient MLEs. If the data are ignorably missing, i.e. the probability of not observing a value does not depend on the unobserved value, then efficient MLEs can be found by maximizing the marginal likelihood, obtained by integrating the likelihood function of the complete data with respect to the missing values. If, otherwise, the data are non-ignorably missing, then a missing value is informative of the unobserved value and a more complex likelihood function must be maximized, obtained by specifying the joint distribution of the complete data and the missing pattern for each subject (Alfò & Maruotti, 2009; Spagnoli et al., 2011; Bartolucci & Farcomeni, 2015).
Motivated by a study of cognitive impairment, this paper describes a HMM that accounts for time-varying latent heterogeneity of both the outcome and the missing value mechanism, by assuming that the multivariate outcome and the missing pattern are conditionally independent given the states visited by a non-homogeneous Markov chain, whose initial and transition probabilities depend on the available covariates.

2 The Chinese Longevity and Health Longitudinal Survey

The data that motivated this paper are drawn from the The Chinese Longevity and Health Longitudinal Survey (CLHLS) and they are available from the National Archive of Computerized Data on Aging. The study includes 358 subjects aged between 80 and 106 in 1998 and followed-up in four subsequent waves in 2000, 2002, 2005, 2008. The outcomes of interest are the answers given by each subject to the 23 items of the Chinese Mini Mental State Examination (MMSE) questionnaire. The scores on each item are binary (e.g. 1 for a correct answer and 0 otherwise). Respondents were asked 5 orientation-related questions, 12 language-related questions, 5 calculation questions and a single-item drawing question. Only 57% of the questionnaires are complete. Partial questionnaires with a number of missing items between 1 and 10 missing items were about the 30% of the sample. The rest of the sample left more than 10 items unanswered. Under this setting, missing values are informative of cognitive functioning and, as a result, the missing value mechanism is not ignorable.

The study further includes a number of individual covariates such as gender, type of residence (rural or urban), whether the subject is sedentary or active and limits in activities of daily living (ADL; six activities including bathing, dressing, eating, indoor transferring, toileting and continence), categorized into three levels: no, one, two or more limits.

3 A hidden Markov model for incomplete multivariate longitudinal data

The data that motivated this study are in the form of incomplete, longitudinal, multivariate binary data, observed on $n$ subjects, at $T$ occasions, with respect to $J$ items, say an array $y_{itj}$, $i = 1 \ldots n, t = 1 \ldots T, j = 1 \ldots J$, where $y_{itj} = 1$ if subject $i$ answered correctly to item $j$ at occasion $t$, and 0 otherwise. Further, let $m_{itj}$ be a binary variable that is equal to 1 if $y_{itj}$ is missing and 0 otherwise. We assume, without loss of generality, that the items are clustered within $G$
groups of $J_1 \ldots J_g \ldots J_G$ questions, $\sum_{g=1}^{G} J_g = J$, and introduce a $G \times J$ matrix $B$, whose generic element $b_{gj}$ is equal to 1 if item $j$ belongs to group $g$ and 0 otherwise. We accordingly refer to $z_{tg} = \sum_{j=1}^{J} y_{tg} b_{gj}$ as the censored number of correct answers in group $g$ and to $m_{tg} = \sum_{j=1}^{J} m_{tg} b_{gj}$ as the number of unanswered questions in group $g$.

The proposed HMM is based on the assumption that, for each subject $i$ at occasion $t$, both the outcomes and the missing indicators are conditionally independent given the level of a latent factor, which can be conveniently described by a multinomial variable $\xi_{it} = (\xi_{it1} \ldots \xi_{itK})$ with one trial and $K$ classes, where $\xi_{itk} = 1$ if the factor takes level $k$ for subject $i$ at time $t$ and 0 otherwise. Under this setting, each subject is associated with a multinomial time series $\xi_i = (\xi_{i0}, \ldots, \xi_{iT})$, which indicates the temporal evolution of the latent factor during the follow-up. By interpreting unobserved factor levels as latent classes, $\xi_i$ can be viewed as a joint segmentation of the longitudinal profile $y_i$ and the missing patterns $m_i$, where each binary component $\xi_{itk}$ indicates the class membership of the pair $(y_{it}, m_{it})$. We assume that the $n$ segmentations $\xi_1 \ldots \xi_n$ are independent multivariate samples, drawn from the joint distribution of a nonhomogeneous Markov chain with $K$ states. Conditionally on the segmentation $\xi_i$, we assume that the distribution of $(y_{it}, m_{it})$ is given by

$$p(y_{it}, m_{it} | \xi_i) = \prod_{g=1}^{G} p^{\text{obs}}_g(\xi_{it}) z_{itg} (1 - p^{\text{obs}}_g(\xi_{it}))^{J_g - z_{itg}} p^{\text{mis}}_g(\xi_{it}) m_{itg} (1 - p^{\text{mis}}_g(\xi_{it}))^{J_g - m_{itg}}$$

where

$$p^{\text{obs}}_g(\xi_{it}) = \prod_{k=1}^{K} \left( p^{\text{obs}}_g \xi_{itk} \right)$$

$$p^{\text{mis}}_g(\xi_{it}) = \prod_{k=1}^{K} \left( p^{\text{mis}}_g \xi_{itk} \right)$$

indicate the group-specific probabilities, which depend on the latent state of subject $i$ at time $t$. Finally the latent trajectories $\xi_i$, $i = 1 \ldots n$ are assumed as independently sampled from a non-homogeneous Markov chain with initial and transition probabilities that are respectively modelled as

$$\log \frac{P(\xi_{itk} = 1)}{P(\xi_{it0} = 1)} = x_{itk}^T \beta_k$$

$$\log \frac{P(\xi_{itk} = 1 | \xi_{i,t-1,h} = 1)}{P(\xi_{ith} = 1 | \xi_{i,t-1,h} = 1)} = x_{it}^T \gamma_{kh}$$
where \( x_i^T \) is the covariate profile of subject \( i \) at time \( t \), whereas \( \beta \) and \( \gamma \) are regression parameters that capture the effect of these covariates on both the initial and the transition probabilities of the latent chain.

4 Results

Examining the CLHLS data by the proposed HMM sheds new light on the factors that influence the cognitive impairment in Chinese adults, extending the cross-sectional results found by Lagona and Zhang (2010) to a longitudinal setting. Impairment in orientation tend to occur later than impairment in calculation, language and drawing. Gender differences in the ability to cope with the questionnaires are time-varying. Only under specific states of the subject, the factors that negatively influence cognitive functioning increase the probability of leaving an item unanswered.

References


