Statistical Analysis and Data Mining: The ASA Data Science Journal Explore this journal >

Volume 10, Issue 1
February 2017
Pages 29–39

View issue TOC

Special Issue:
CLADAG 2015 Special Issue: Selected Papers on Classification and Data Analysis

Original Article

Latent Markov and growth mixture models for ordinal individual responses with covariates: A comparison

Fulvia Pennoni, Isabella Romeo

First published:
5 January 2017 Full publication history

DOI:
10.1002/sam.11335 View/save citation

Cited by (CrossRef):
0 articles Check for updates Citation tools

Am score 3

Funding Information

Abstract

Objective

We review two alternative ways of modeling stability and change of longitudinal data by using time-fixed and time-varying covariates for the observed individuals. Both the methods build on the foundation of finite mixture models, and are commonly applied in many fields but they look at the data from different perspectives. Our attempt is to make comparisons when the ordinal nature of the response variable is of interest.

Methods

The latent Markov model is based on time-varying latent variables to explain the observable behavior of the individuals. It is proposed in a semiparametric formulation as the latent process has a discrete distribution and is characterized by a Markov structure. The growth mixture model is based on a latent categorical variable that accounts for the unobserved heterogeneity in the observed trajectories and on a mixture of Gaussian random variables to account for the variability in the growth factors. We refer to a real data example on self-reported health status to illustrate their peculiarities and differences.

1 INTRODUCTION

The analysis of longitudinal or panel data by using latent variable models has a long and rich history mainly in the social sciences. In the past several decades, the increased availability of large and complex data sets, have witnessed a sharp increase in interest in this topic. Nowadays, it demands the development of increasingly rigorous statistical analytic methods that can be proved useful for data reduction as well as for inference. Among the different proposals available there are two main broad classes of models: one tailored to consider the transition over time and the other focused on growth or trajectory analysis. Among the former, we discuss the latent Markov (LM) model which is mainly used for the analysis of categorical data. Among the second class, the growth mixture model (GMM) is originally employed with observed continuous response variables. In the following we compare the models to account for the recent improvements proposed in literature. Previous comparisons can be found in [1, 2] and some hints are available in [3]. We consider measurements on an ordinal scale to illustrate similarities and differences between these models.

The LM models may be classified as observation-driven models tailored for many types of longitudinal categorical data as showed recently in [4, 5]. The evolution of the individual characteristics of interest over time is represented by a latent process with state occupation probabilities that are time-varying. They are extensions of the latent class model [6] when multiple occasion of measurements are available and of Markov chain models for stochastic processes when an error term is included in the observations. They allow for unobserved heterogeneity among individuals or within the latent states. Even if the first basic model formulation proposed by Wiggins [7] does not include the covariates, at present time-constant and time-varying covariates can be added in the measurement or in the latent part of the model. Wiggins proposed this model at Columbia in a social science research project when Paul Lazarsfeld was principal investigator (see for more details http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/lazarsfeld-paul-f.pdf). In 1955 in his Ph.D. dissertation he analyzed the applicative example of a single item of human behavior moving over time in a nonexperimental context. When the model is formulated according to a discrete time-dependent latent process it may be classified as a semiparametric approach. It allows modeling with different data in applications in fields such as medicine, sociology, biology, or engineering (see also [8, 9]). Some of the connections with the hidden Markov model employed to analyze time-series data are illustrated in [10]. The hidden Markov model was also developed in the social science field to study sudden changes in learning processes by Miller [11]. An alternative model formulation to assess causal effects under the potential outcome framework [12] has been recently proposed in [13].

Conventional growth models or growth curve models (GCMs) are viewed either as hierarchical linear models or as structural equation models. Their use in analyzing continuous response variables has been widely discussed in the literature (see, among others [14, 15]). Their use in modeling and analyzing categorical data has recently received more attention [16, 17]. Latent growth modeling was first proposed independently in [18, 19] in relation to the longitudinal factor analysis and later extended and refined in [20-22]; see also [23].

The GCM aims at studying the evolution of a latent individual characteristic in order to estimate the trajectories by accounting for individual variability about a mean population trend. It imposes a homogeneity assumption, requiring that all individuals follow similar trajectories. The GMM proposed by [24] (see also [25, 26]) is a generalization of the GCM which accounts for the heterogeneity in the observed development trajectories by employing a latent categorical variable. The finite mixture of linear and multinomial regression models allows us to disentangle the between-individual differences and the within-individual pattern of changes through time (see also [27, 28]). It is a parametric approach where the population variability in growth is modeled by a mixture of subpopulations with different Gaussian distributions.

A specific case of the GMM is the latent-class growth curve model (LGCM) (see, among others, [29-31]), also termed as latent class regression model by [32]. Another terminology employed in [33] is latent class growth analysis (LCGA). The multinomial model is used to identify the homogeneous groups of developmental trajectories by avoiding the random effects of Gaussian distribution assumption. The individuals in each class share a common trajectory [34] without considering the between-class heterogeneity. Therefore, in the LGCM, the individual heterogeneity is captured completely by the mean growth trajectories of the latent classes. However GMM allows us to model the class-specific variance components (intercept and slope variance). For a more complete comparison between GMM and LGCM, see also [35]. An alternative extension of these models to the counterfactual context has been proposed in [36].

We illustrate two recent extensions of the LM model and GMM where the ordinal response is made by thresholds imposed on an underlying continuous latent response variable. We show how the discrete support for the latent variable used in the LM model framework can be appropriate in this context. The models are compared on how they allow covariates, how they make inference, on their computational features required to achieve the estimates, and on their ability to classify units and their predictive power. Our proposal to compare them in terms of fitting, parsimony, interpretation, and prediction is an attempt to review the recent literature on these models for panel data. The results of the model fitting are illustrating through a data set on longitudinal study aimed at describing self-perceived health status, which also appears in other published scientific articles (see, among others [37]).

The structure of the paper is as follows. In Section 2 we introduce the basic notation for both models and we summarize the main features concerning the estimation issues. In Section 3 we demonstrate the effectiveness of the models explaining their purposes in relation to the applied example and their results. In the last section we draw some concluding remarks.

2 MAIN NOTATION AND ILLUSTRATION OF THE MODELS

One way to afford the issue of ordinal response variables consists in deriving a conditional probability model from a linear model for a latent response variable. The observed variables are obtained by categorizing the latent continuous response that may be related, for example, to the amount of understanding, attitude, or wellbeing required to respond in a certain category. Let Y_{it} be the observed ordinal variable for individual i, for $i = 1, \ldots, n$ at time t, $t = 1, \ldots, T$. We assume an underlying continuous latent variable Y^*_{it}, via a threshold model given by

$$Y_{it} = s \quad \text{iff} \quad \tau_{s-1} < Y^*_{it} \le \tau_s,$$

where $s = 1,2,\ldots, S$ and $-\infty = \tau_0 < \tau_1 < \tau_2 <$ □ $< \tau_{s-1} < \tau_s = +\infty$ are the cut-off points by which it is possible to achieve a unique correspondence. With S response categories, there are $S - 1$ threshold parameters, τ_s, $s = 1,2,\ldots, S - 1$.

2.1 LM models for ordinal data

Under the basic model we assume the existence of a discrete latent process such that

$$Y_{it}^{*} = \alpha_{it} + \varepsilon_{it},$$

with $\alpha_{i1},\ldots, \alpha_{iT}$ following a hidden Markov chain with state space $\xi_1,\ldots, \xi_k$, initial $\pi_u = p(\alpha_{i1} = \xi_u)$, and transition probabilities $\pi_{u|\bar{u}} = p(\alpha_{it} = \xi_u | \alpha_{i,t-1} = \xi_{\bar{u}})$, $\bar{u}, u = 1,\ldots, k$. Moreover, $_{it}$ is a random error with normal or logistic distribution.

In the case of time-varying or time-fixed covariates collected in the column vectors x_{it}, the model is extended as:

$$Y_{it}^{*} = \alpha_{it} + \boldsymbol{x}_{it}'\beta + \varepsilon_{it},$$

so as to include these covariates in the measurement model concerning the conditional distribution of the response variables given in the latent process. The covariates are allowed in the latent part of the model; however, the model is better identified when the covariates are stored in the latent or in the measurement model. The choice is related to the research question and the aims of the analysis.

The model has a simple structure if the discrete latent process follows a first-order homogeneous Markov chain and we can assume the conditional independence of an observed response variable Y_{it} in relation to the other responses given the latent process for $i = 1,\ldots, n$, $t = 1,\ldots, T$. This is called the local independence assumption. The conditional distribution of the responses is denoted by $f_t(y | u, x)$, $u = 1,\ldots, k$, whereas the latent stochastic process U has initial probability function $p(u)$, for $u = 1,\ldots, k$, and transition probability function $p_t(u | \bar{u})$, where $t = 2,\ldots, T$, $u, \bar{u} = 1,\ldots, k$, and k denote the discrete number of latent states. Therefore, a semiparametric model results. A generalized linear model parameterization [38] allows us to include properly the covariates in the measurement model. In this way, by using suitable link functions we can allow for specific constraints of interest and we can also reduce the number of parameters.

An effective way to include the covariates in the measurement model is to consider

$$\boldsymbol{\eta}_{tux} = \boldsymbol{C}\log[\boldsymbol{M}\boldsymbol{f}_t(u,\boldsymbol{x})],$$

where C is a suitable matrix of contrasts, M is a marginalization matrix with elements 0 and 1, which sums the probabilities of the appropriate cells and the operator log is coordinate wise, $f_t(u, x)$ is a c-dimensional column vector with elements $f_t(y | u, x)$ for all possible values of y. In the following, $\eta_{ty|ux}$ denotes each element of η_{tux} where $y = 1,\ldots, s - 1$. Within this formulation, we can state some hypothesis of interest by constraining the model parameters according to the research question related to the application. For example, an interesting formulation is the following:

$$\eta_{y|ux} = \beta_{1y} + \beta_{2u} + \boldsymbol{x}'\boldsymbol{\beta}_3, \quad y = 1, \ldots, s-1, u = 1, \ldots, k, \tag{1}$$

where the levels of β_{1y} are cut-off points or threshold parameters, β_{2u} are intercepts specific to the corresponding latent state, and β_3 is a vector of parameters for the covariates. The above is possible once we define the global logits [38] on the conditional response mass function:

$$\eta_{y|ux} = \log\frac{f(y|u,\boldsymbol{x}) + \cdots + f(s-1|u,\boldsymbol{x})}{f(0|u,\boldsymbol{x}) + \cdots + f(y-1|u,\boldsymbol{x})}, \quad y = 1, \ldots, s-1.$$

We carry out the estimation of the model parameters in two ways: by using the maximum likelihood method through the EM algorithm [39] or by the Bayesian methods applying the Markov Chain Monte Carlo methods [40]. Within the first choice, the log-likelihood is maximized according to the following steps until convergence:

E. step to compute the expected value of the complete data log-likelihood given the observed data and the current value of θ, which denotes all the model parameters;

M. step to maximize this expected value with respect to θ and thus update θ.

We use the recursions developed in the hidden Markov literature by [41] and by [42] to compute the quantities of interests. They enable computing efficiently the expected values of the random variables involved in the complete data log-likelihood:

$$\begin{aligned}\ell^{*}(\boldsymbol{\theta}) = &\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{x}\sum_{y=0}^{s-1} a_{tuxy}\log f_t(y\,|u,\boldsymbol{x}) + \sum_{u=1}^{k} b_{1u}\log p(u)\\ &+ \sum_{t=2}^{T}\sum_{\bar{u}=1}^{k}\sum_{u=1}^{k} b_{tu\bar{u}}\log p_t(u|\bar{u}),\end{aligned}$$

where a_{tuxy} is the number of individuals that are in latent state u and provide response y at occasion t, b_{1u} is the frequency of the latent state u, and $b_{tu\bar{u}}$ is the number of transitions from state $\bar{u}$ to state u at occasion t.

As for other mixture models [43] there may be many local optima, therefore the estimation is carried out by considering multiple sets of starting values for the chosen algorithm. A drawback of the EM algorithm is that it does not provide a direct quantity to assess the precision of the maximum likelihood estimates. It is possible to consider the missing information principle. In the case of the regular exponential family [44], the observed information is equal to the complete information minus the missing information due to the unobserved components [45, 46]. For an implementation of the above and for the directed acyclic Gaussian graphical models with hidden variables see [47]. Its computational burden is low over that required by the maximum likelihood estimation.

The model selection may be based on a likelihood ratio (LR) test statistics between the model with k latent classes and that with $k+1$ latent classes for increasing values of k, until the test is not rejected. However, we need to employ the bootstrap to obtain a p-value for the LR test. It is based on a suitable number of samples simulated from the estimated model with k latent classes [48]. In [49] they select the best parsimonious model through a consistent estimator based on the parametric bootstrap. The best model is one among those with the proposed number of latent classes.

We select the number of latent states according to the information criteria most commonly employed: the Akaike information criterion (AIC, [50]) and the Bayesian information criterion (BIC, [51]). We recall that, when the states are selected according to the model with the smallest value of BIC, we decrease the maximum of the log-likelihood value, considering also the total number of individuals. Their performance has been studied in-depth in the literature on mixture models (see, among others [43], Chapter 6). They are also employed in the hidden Markov literature for time-series, where they are penalized by the number of time occasions (see, among others [52]). The BIC is usually preferred to AIC, as the latter tends to overestimate the number of latent states but it may be too strict in certain cases (see, among others [53]). The theoretical properties of BIC in the LM models framework are still not well established. However, BIC is a commonly accepted choice criterion for these models as well as to choose the number of latent classes for the latent class model (see, among others [54]). In [5], this criterion is also used together with other diagnostic statistics measuring the goodness-of-classification. A more recent study [55] compares the performance of some likelihood and classification-based criteria, such as an entropy measure, for selecting the number of latent states when a multivariate LM model is fitted to the data.

An interesting feature of the LM model concerns prediction. As shown in [5] the local decoding allows prediction of the latent state for each individual at each time occasion by maximizing the estimated posterior function of the latent process. The global decoding employing the Viterbi algorithm [56], (see also [57]) allows us to obtain the most a posteriori likely predicted sequence of states for each individual. The joint conditional probability of the latent states given the responses, and the covariates $\hat{f}_{U|X,Y}(u\,|\boldsymbol{x},\boldsymbol{y})$ are computed by using a forward recursion according to the maximum likelihood estimates of the model parameters, where u denotes a configuration of the latent states. The optimal predicted state

$$\hat{u}_t^* = \arg\max_u \hat{r}_t(u)\hat{p}_{t+1}(u|\hat{u}^*_{(t-1)})$$

is found by considering $\hat{r}_1(u) = \hat{p}(u|x)\prod_t \hat{f}_1(y_1|u_1,x)$, where the hat denotes the value of the parameter at the maximum of the log-likelihood of the model of interest, for $u = 1, \ldots, k$; and computing in a similar way $\hat{r}_t(\bar{u})$, for $t = 2, \ldots, T$ and $\bar{u} = 2, \ldots, k$; then maximizing such that $\hat{u}_T^* = \arg\max_u \hat{r}_T(u)$.

2.2 Growth mixture models

The GCMs provide the estimated shapes of the individual trajectories accounting for within and between individual differences. The measurement model concerning the observed responses deals with individual growth factors. The latent model is related to the means, variances, and covariances of the growth factors to explain between-individual differences. First we recall the LGCM and then the GMM. The LGCM without covariates is defined by the following equations:

$$\begin{aligned} Y_{it}^* &= \alpha_i + \lambda_t\beta_i + \lambda_t^2 q_i + \varepsilon_{it}, \\ \alpha_i &= \mu_\alpha + \zeta_{\alpha i}, \\ \beta_i &= \mu_\beta + \zeta_{\beta i}, \\ q_i &= \mu_q + \zeta_{q_i}, \end{aligned} \tag{2}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where α_i and β_i are named intercept and slope growth factor respectively, and q_i is the quadratic growth factor. To allow identifiability, the coefficient of the intercept growth factor is fixed to 1. Therefore, it equally influences the repeated measures across the waves and it remains constant across time for each individual. Different values can be assigned to the coefficient λ_t related to each time occasion t, in order to dispose of growth curves with different shapes that are linearly or not linearly dependent on time. In order to define a growth model with equidistant time points, the time scores for the slope growth factor are fixed at $0, 1, 2, \ldots, T-1$ (see, among others [15]). The first time score is fixed at zero and the intercept growth factor can be interpreted as the expected response at the first time point. The time scores for the quadratic growth factor are fixed at $0, 1, 4, \ldots, (T-1)^2$ to allow for a quadratic shape of the trajectory, and for a linear growth model the quadratic growth factor q_i is fixed at 0 for all i, $i = 1, \ldots, n$.

The measurement errors $_{it}$ in Equation (2) are not correlated across time, they are i.i.d. disturbances. Because there is no intercept term in the measurement model, the mean structure of the repeated measures is determined entirely by means of the latent trajectory factors. In the structural model, the parameters μ_α, μ_β, and μ_q are the population means of the intercept, slope, and the quadratic term respectively; $\zeta_{\alpha i}$ is the deviation of α_i from the population mean intercept, $\zeta_{\beta i}$ is the deviation of β_i from the population mean slope, and ζ_{q_i} is the corresponding deviation from the population mean quadratic factor. They are assumed to follow a multivariate Gaussian distribution with zero means and variances denoted by $\psi_{\alpha\alpha}$, $\psi_{\beta\beta}$, and ψ_{qq} respectively and they are uncorrelated with $_{it}$. The covariance of the intercept and the slope growth factor is $\psi_{\alpha\beta}$, those of the quadratic factor with the intercept and the growth factor are $\psi_{\alpha q}$ and $\psi_{\beta q}$, respectively. When the response is ordinal or categorical, the thresholds are assumed to be equal for each measurement occasion by imposing the constraint $\tau_{st} = \tau_s$ for all t, $t = 1, \ldots, T$ and the constraint $\mu_\alpha = 0$ is also required.

In the conditional growth model, the time-fixed covariates are included as predictors of the growth factors or as direct predictors of the response variable. Time-varying covariates can only be included as predictors in the measurement model according to the following equations where the quadratic term as in Equation (2) is deleted to simplify the notation:

$$\begin{aligned} Y^*_{it} &= \alpha_i + \lambda_t\beta_i + \omega_{it}\gamma_t + \varepsilon_{it}, \\ \alpha_i &= \mu_\alpha + x_i'\gamma_\alpha + \zeta_{\alpha i}, \\ \beta_i &= \mu_\beta + x_i'\gamma_\beta + \zeta_{\beta i}, \end{aligned} \tag{3}$$

for $i = 1, \ldots, T$ and $t = 1, \ldots, T$, where γ_α and γ_β are vectors of parameters for the time-fixed covariates x_i on α_i and β_i, respectively, and γ_t is the vector of parameters for the time-varying covariates ω_{it} on the measurement model.

The unconditional GMM is defined by a latent categorical variable U accounting for the unobserved heterogeneity in the development among individuals. It represents a mixture of subpopulations whose membership is inferred by the data (for a review, see, among others [15, 58]). It is characterized by the following equations:

$$\begin{aligned} Y^*_t &= \sum_{u=1}^{k} p_u(\alpha_u + \lambda_{tu}\beta_u + \varepsilon_{tu}), \\ \alpha_u &= \mu_{\alpha u} + x'\gamma_{\alpha u} + \zeta_{\alpha u}, \\ \beta_u &= \mu_{\beta u} + x'\gamma_{\beta u} + \zeta_{\beta u}, \end{aligned}$$

for $t = 1, \ldots T$, where p_u is the probability of belonging to latent class u, for $u = 1, \ldots, k$ which defines the latent trajectory, with the constraints $p_u \geq 0$ and $\sum_{u=1}^{k} p_u = 1$, where k is equal to the number of mixture components. The thresholds τ_s are unknown and they are estimated and constrained to be equal across time and latent classes. The intercepts of the growth factors may vary across latent classes. With categorical response variables, the growth factor referred to the last class is constrained to zero for identifiability issues and the others are estimated from the model. The variances and covariance of the growth factors can be allowed to be class-specific or constrained to be equal. Residuals of the growth factors and of the measurement model are assumed with a Gaussian distribution within each latent class. As in Equation (3) only time-fixed covariates may be included to infer the latent class through a multinomial logistic regression model since the latent variable is typically viewed as time invariant. Therefore, the GMM reduces to the GCM when $k = 1$ and to the LGCM when the within-class growth factor variance and covariances $\psi_{\alpha u}$, $\psi_{\beta u}$, $\psi_{\alpha\beta u}$ are set to zero for all $u = 1, \ldots, k$. In the latter case, the between-individual variability is captured only by the latent class membership. The thresholds are estimated with the mean cumulative response probabilities for a specific response category at each measurement occasion by the estimated distribution of the latent growth factors.

The maximum likelihood estimation of the model parameters when there are categorical response variables and continuous latent variables requires numerical methods. The computation is carried out by using Monte Carlo integration [15, 59]. As in the standard Gaussian mixture models, imposing constraints on the covariance matrices of the latent classes ensures the absence of singularities and potentially reduces the number of local solutions [24, 28]. The model selection concerns the choice of the number of the latent classes and the order of the polynomial of the group's trajectories. The most common applied empirical procedure is the following: first the order of the polynomial is assessed by estimating both linear and nonlinear unconditional GCM, or GMM with $k = 1$, GMM(1) in the following. Then, the number of latent classes is determined according to the unconditional model in order to avoid an over-extraction of the latent classes (see also [60]). Finally, the covariates are added in the model as predictors of the latent classes.

The LR statistic is employed for the model selection also by considering the bootstrap (see, among others [61]) as illustrated in the previous section. The number of latent classes is selected according to the AIC or BIC indices illustrated in Section 2.1. The relative entropy measure [62] is commonly employed to state the goodness of classification:

$$E_k = 1 - \frac{\sum_{i=1}^{n}\sum_{u=1}^{k} -\widehat{p}_{iu}\log(\widehat{p}_{iu})}{n\log(k)}, \tag{4}$$

where $\widehat{p}_{iu}$ is the estimated posterior probability of belonging to the u-th latent class at convergence, k is the number of latent classes, and n is the sample size. The values approach 1 when the latent classes are well separated. However, we notice that it differs from the normalized entropy criterion defined by [63] which instead divides the first term of the Equation (4) by the difference between the log-likelihood of the model with k classes and the one with just one class. The above criteria may lead to a model lacking of interpretability in terms of latent classes or in which only few individuals are allocated in a class. As suggested by many authors such a choice needs also to be guided by the research question as well as by theoretical justification and interpretability [64-66]. The optimal number of classes derived from the LGCM is always bigger than the optimal number of classes derived from GMM. Within the LGCM, individuals with slightly different growth parameters are allocated to a different latent class compared with the GMM (see, among others [67]).

3 REAL DATA EXAMPLE: THE HEALTH AND RETIREMENT STUDY

In order to show the main differences among the models illustrated in the previous section, we consider a longitudinal study aimed at describing self-perceived health status. The latter is a frequently used way to establish health policy and care as the repeated subjective health assessment reflects the self-perception of health and how it is going to evolve over time. It is recorded by one item with response categories defined according to an ordinal variable. The data is taken from version I of the RAND HRS data, collected by the University of Michigan (see also http://www.cpc.unc.edu/projects/rlms-hse and http://www.hse.ru/org/hse/rlms). The 30 406 respondents were asked to express opinions on their health status at $T = 8$ approximately equally spaced occasions, from 1992 to 2006. After considering only individuals with no missing data, we ended up with a sample of $n = 7074$ individuals. The response variable is measured on a scale based on five categories: “poor”, “fair”, “good”, “very good”, and “excellent”. For each individual, some covariates are also available: gender, race, education, and age (at each time occasion). The study relies on the investigation of the population heterogeneity in the health status perception as well as on prediction of features needs to be especially tailored for those elders who are identified to share the most difficult health conditions.

First, we summarize the estimation process for both models presented in Section 2 and then we make some comparisons on the estimated quantities. The estimation of the LM models is undertaken in the R environment [68] through the library LMest (V2.2) [69] that is available on the Comprehensive R Archive Network. This version also accounts for the covariates on the latent part of the model and missing values on the responses. The estimation of the growth models is undertaken via the commercial software MPLUS (V7.2). The syntax code is available from the authors upon request.

For the LM model parameterized as in Equation (1) we employ the model search procedure as illustrated in Section 2.1 to find the best model among those with a number of latent states from 1 up to 11. The search strategy which is implemented to account for the multimodality of the likelihood function is based on estimating the same model many times with the same number of states by using deterministic and random starting values for the EM algorithm. The number of different random starting values is proportional to the number of latent states. The relative log-likelihood difference is evaluated by considering a tolerance level equal to 10^{-8}. The model is estimated for an increasing number of latent states while checking for the replication of likelihood values. The best model is the one with nine latent states according to the BIC values as showed in Table 1 denoted by LM(9) in the following. The table also reports the AIC values and the number of free parameters.

Table 1. Fitted statistics for an increasing number of latent states from 1 to 11 of the LM model with covariates and number of parameters

	Log-likelihood	AIC	BIC	#par
LM(1)	−80 623.52	161 267.0	161 335.7	10
LM(2)	−69 789.21	139 604.4	139 693.6	13
LM(3)	−65 707.82	131 451.6	131 575.2	18
LM(4)	−63 968.06	127 986.1	128 157.7	25
LM(5)	−63 293.98	126 656.0	126 889.3	34
LM(6)	−63 062.23	126 214.5	126 523.4	45
LM(7)	−62 894.29	125 904.6	126 302.7	58
LM(8)	−62 739.12	125 624.2	126 125.3	73
LM(9)	−62 645.69	125 471.4	126 089.1	90
LM(10)	−62 615.99	125 450.0	126 198.2	109
LM(11)	−62 650.58	125 561.2	126 453.5	130

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion; LM, latent Markov; #par, number of parameters.

The estimated cut-off points of the LM(9) model are $\hat{\tau}_1 = 8.261, \hat{\tau}_2 = 4.559, \hat{\tau}_3 = 0.800, \hat{\tau}_4 = -3.470$. The estimated initial probabilities are reported in Table 2 together with the support points. The estimated support points are arranged in increasing order, in order to interpret the resulting latent states from the worst (latent state 1) to the best (latent state 9) health conditions. We notice from Table 2 that 11% and 19% of individuals are in the second and third latent states respectively, which are worse states with respect to latent states 6 and 8. Table 4 reports the matrix of the estimated transition probabilities between latent states. The only greater probabilities than 0.10 in the elements adjacent to the diagonal are those of the transition from the first to the second latent state and from the second to the third. For the latent state 4, the probability to move to the latent states 7 or 8 or 9 is higher than 0.10. They show that the individuals belonging to this state, perceiving bad health conditions at the beginning of the survey, have some probability to feel better (to improve their health conditions) over time. For the latent state 8, the probability of moving to latent state 3 or 4 or 5 are higher than 0.10.

Table 2. Estimated support points and parameters referring to the initial probabilities of the chain of the LM(9) model

Latent state	Support points	Initial probabilities
1	−8.657	0.047
2	−4.941	0.117
3	−2.456	0.192
4	−1.147	0.028
5	−0.224	0.213
6	2.062	0.189
7	4.303	0.121
8	5.159	0.213
9	7.357	0.067

Abbreviation: LM, latent Markov.

Table 3 shows the effect of the covariates on the probability of reporting a certain level of the health status. In particular, women tend to report worse health status than men (the odds ratio for females versus males is equal to (exp(−0.185) = 0.831), whereas white individuals have a higher probability of reporting a good health status with respect to non-whites (the odds ratio for non-whites versus whites is equal to (exp(−1.341) = 0.261). We also observe that better educated individuals tend to have a better opinion about their health status especially those with a high educational qualification. Finally, the effect of age is decreasing over time and its trend is linear as the quadratic term of age is not significant.

Table 3. Estimates of the vector of the regression parameters of the LM(9) model

Coefficient	Female	Non-white	Some college	College and above	Age	Age^2
β	−0.185	−1.341	1.37	2.461	−0.125	−0.001
se	0.075	0.109	0.092	0.104	0.007	0.026

Abbreviations: LM, latent Markov; se, standard errors.

Table 4. Estimates of the transition probabilities under the LM(9) model (probabilities out of the diagonal greater than 0.1 are in bold)

	$\hat{\pi}_{u\|\bar{u}}$								
	1	2	3	4	5	6	7	8	9
1	0.796	**0.182**	0.000	0.001	0.006	0.001	0.002	0.012	0.000
2	0.053	0.822	**0.106**	0.002	0.000	0.000	0.000	0.017	0.000
3	0.008	0.020	0.868	0.004	0.061	0.001	0.000	0.038	0.000
4	0.026	0.013	0.001	0.336	0.006	0.039	**0.155**	**0.292**	**0.132**
5	0.002	0.024	0.015	0.000	0.887	0.066	0.006	0.000	0.000
6	0.000	0.004	0.024	0.003	0.024	0.896	0.045	0.001	0.003
7	0.001	0.004	0.001	0.052	0.025	0.009	0.845	0.001	0.062
8	0.018	0.061	**0.189**	**0.301**	**0.153**	0.000	0.000	0.278	0.000
9	0.000	0.000	0.000	0.050	0.006	0.051	0.072	0.000	0.821

Abbreviation: LM, latent Markov.

In Figure 1 we compare the individual response profiles of the LM(9) model obtained by using the estimated posterior probabilities according to the rules illustrated in Section 2.1. They are related to the white female participants over 65 years of age at the third wave of interview, who are highly educated. They may constitute a special group of people to account for. From Figure 1 we notice that some profiles are less regular than others: they detect those females whose health status may strongly decline due to events that are not observed through the covariates.

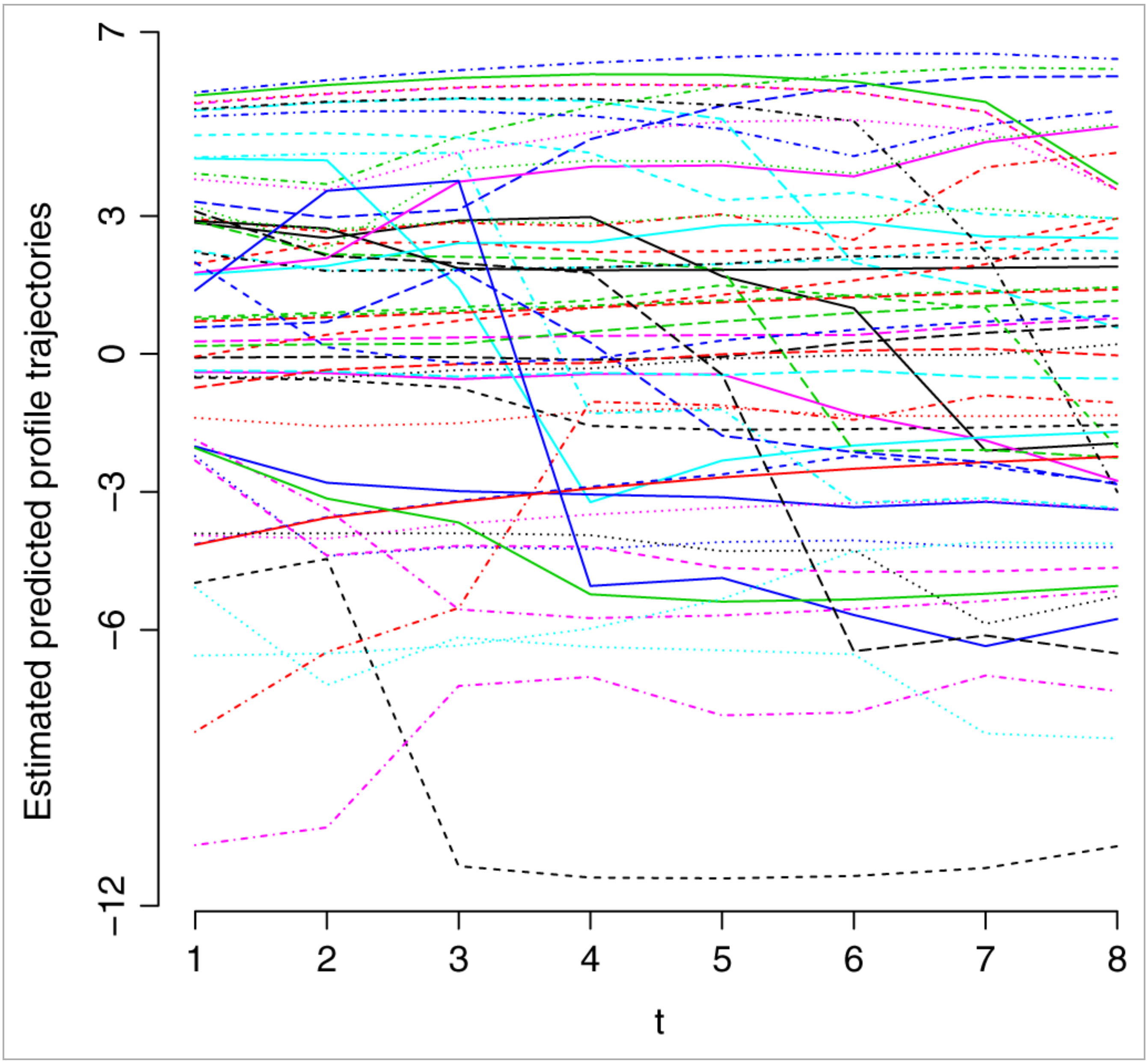

Figure 1.

Open in figure viewer

Individual profiles for a selected group of individuals for the LM(9) model. LM, latent Markov.

For the growth models, we detect the best model within the class of GMMs according to the model strategy illustrated at the end of Section 2. As the first step, we estimate two GMMs without covariates with just one latent class in which the respondents' opinions about their health are specified as a function of linear and nonlinear growth patterns. The GMM with a quadratic effect shows a log-likelihood equal to −63 996.8 and the BIC index equal to 128 100 with 12 parameters. This model is preferred according to a BIC index as the GMM without the quadratic effect results in the log-likelihood equal to −63 116.3 and the BIC value equal to 128 303.5 with eight parameters (the χ^2 test is equal to 1761 with four degrees of freedom which is significant). As the second step, we reject the hypothesis of homogeneity within groups since the log-likelihood of the linear model under this assumption decreases to −83 152.7. When we consider the quadratic term we reach three dimensions of integration, the computer burden increases exponentially and the model with a high number of latent classes does not reach the convergence. The estimated parameters of the linear GMM model denote that the perception of a good health status decreases over time. The variances of the intercept and of the slope factor are significant, indicating the existence of individual differences in growth trajectories. As a third step, we fit the selected GMM model without covariates by considering the existence of a mixture of Gaussian distributions from two up to five components with varying patterns of the growth trajectories.

Table 5 shows the results. We select the model with three latent classes according to the BIC index denoted as GMM(3) as the models with a higher number of components do not reach the convergence criteria. The model with four latent classes has the same log-likelihood value of the model with three latent components. The best log-likelihood value for the model with five latent classes is not replicated with different starting values. As a last step, we include in the model of Equation (3) time-fixed covariates, taken as constants across the latent classes. Their coefficients are significant with the exception of the quadratic effect of age. The resulting model has a log-likelihood equal to −63 421.0 and a BIC index equal to 127 143.3 with 34 parameters. The entropy value as in Equation (4) is equal to 0.763.

Table 5. Selection of the number of latent classes of the GMM without covariates

Latent class	Log-likelihood	BIC	#par	Entropy
1	−64 116.3	128 303.5	8	1.000
2	−64 092.3	128 282.2	11	0.599
3	−63 982.3	128 088.7	14	0.719
4	−63 982.2	128 115.1	17	0.428
5	−63 977.2	128 131.7	20	0.746

Abbreviations: BIC, Bayesian information criterion; GMM, growth mixture model; #par, number of parameters.

The estimated probabilities of GMM(3) and the average conditional probability of belonging to each latent class are displayed in Table 6. This is a common employed way to assess the tenability of the selected model as the average posterior probability of group membership for each trajectory is considered as an approximation of the trajectories' reliability. The posterior probabilities are used to assign each individual membership to the trajectory that best matches. Values of 0.70 or 0.80 are reference values in the literature to group individuals with a similar pattern of change in the same latent class. Table 6 shows the classification probabilities for the selected GMM(3) by considering the most likely latent class membership (row) by the average conditional probabilities (column). We notice that contrary to our expectation, the diagonal values referred to the first and third latent class are lower than that of the second latent class meaning that these classes are not properly identified. The percentage of units belonging to the first and third latent classes according to the estimated posterior probabilities is equal to 10.8% and 3.2%, respectively. From Table 7, the estimated coefficients of the covariates on the growth factor are not high and the sign of the female coefficient is reversed in comparison to that estimated by employing the LM model. Therefore, females tend to report better health status than man. This is probably due to the poor reliability of the selected model. The high education shows the highest positive estimated coefficient on the intercept factor.

Table 6. Classification probabilities for the GMM(3) with covariates according to the most likely latent class membership (row) by the average conditional probabilities (column)

	1	2	3
Class 1	0.436	0.556	0.008
Class 2	0.022	0.973	0.005
Class 3	0.028	0.436	0.537

Abbreviation: GMM, growth mixture model.

Table 7. Estimates of the regression parameters of the intercept and slope growth factor of the GMM(3) with covariates

Coefficient	Female	Non-white	Some college	College and above	Age
γ_α	0.265	−1.506	1.037	1.876	−0.044
se	0.103	0.170	0.136	0.148	0.009
γ_β	0.005	0.032	−0.040	−0.071	0.000
se	0.012	0.015	0.016	0.018	0.001

Abbreviations: GMM, growth mixture model; se, standard errors.

As shown in Table 8 the estimated covariance is negative, meaning that the individuals with the highest values of the intercepts at the first occasion (e.g. with better perceived health) change more rapidly into a worse perception. Figure 2 illustrates the estimated trajectories where the first latent class identifies the individuals with initial poor health status and a slow decline in their health, the second latent class those with a better initial health status and a slightly faster decline compared to the first class and the third latent class individuals perceiving a strong worsening of their health status over time.

Table 8. Estimates of the structural parameters of GMM(3) with covariates

Coefficient	Estimates	se	Coefficient	Estimates	se
$\mu_{\alpha(1)}$	−6.734	0.498	$\mu_{\beta(1)}$	−0.105	0.090
$\mu_{\alpha(2)}$	−2.302	0.443	$\mu_{\beta(2)}$	−0.193	0.069
$\mu_{\alpha(3)}$	0.000	0.000	$\mu_{\beta(3)}$	−1.292	0.118
ψ_α	6.501	0.422	ψ_β	0.065	0.005
$\psi_{\alpha\beta}$	−0.272	0.039			

Abbreviations: GMM, growth mixture model; se, standard errors.

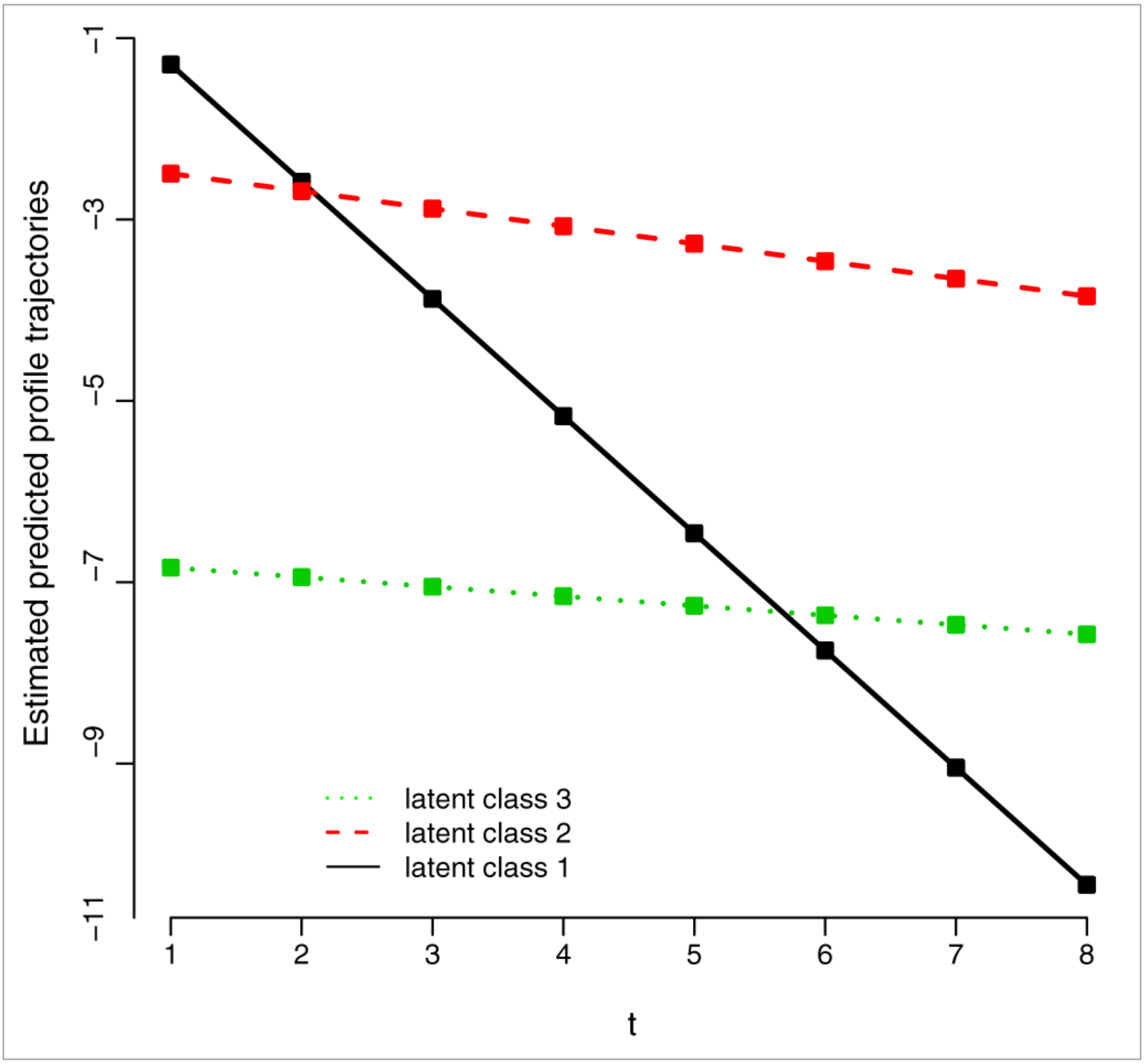

Figure 2.

Open in figure viewer

Response profile plot for the GMM(3) with covariates. GMM, growth mixture model.

4 CONCLUDING REMARKS

We propose a comparison between the LM models and the GMMs when the interest lies in modeling longitudinal ordinal responses and time-fixed and time-varying individual covariates. The interest in this topic is relevant since in many different contexts ordinal data are a way to account for the importance given by an item or to measure something which is not directly observable.

The LM model is a data-driven model which relays on a latent stochastic process following a first-order Markov chain with the fundamental principle to estimate transitions between latent states and to capture the influence of time-varying and time-fixed covariates on the observed transitions. GMM exploits a latent categorical variable to allow the unobserved heterogeneity in observed development trajectories. The latent variable is time invariant and it describes the trend through a polynomial function allowing for time-fixed covariates. We illustrate the main features of the models and their performance by referring to a specific application based on real data in which the ordinal response variable describes the self-perceived health status. The aim is also to estimate a life expectancy for longevity.

We can summarize the main differences between the LM model and the GMM according to the following characteristics: (1) the model estimation and selection procedure leading to the choice of the number of the latent states or classes, (2) the way they relate the conditional probabilities of the responses to the available individual covariates, (3) the model capability to use the posterior probabilities in order to get profiles for each latent class membership. We show that the LM model outperforms the GMM mainly because it is more rigorous on each of the above points. With reference to (1) the model choice is more complex for the GMM and it starts with the model without covariates. We found that the Monte Carlo integration for the GMM with a number of latent classes up to three, leads to improper solutions. The selection of the best model is more straight for the LM model, however it requires a search strategy to properly initialize the EM algorithm and therefore it is computationally demanding when the number of latent states in the model is high. With reference to (2) the covariates are better handled by the LM model since they are allowed according to a suitable parametrization for categorical data such as global logits. While in the LM model the covariates may affect the measurement part of the model or may influence the latent process, in the GMM they can affect both but in the measurement model, only time-fixed covariates are allowed. Then, when the interest is on detecting subpopulations in which individuals may be arranged according to their perceived health status, the LM model is more appropriate. The GMM can be useful when just a mean trend is of interest and the expected subpopulations are not too many. With reference to (3) the predictions of the LM model are based on local and global decoding. The first is based on the maximization of the estimated posterior probability of the latent process and the second on a well-known algorithm developed in the hidden Markov model literature to get the most a posteriori likely predictive sequence. In the GMM, the prediction is based on the maximum posterior probability and as shown in the example it may not be precise when the internal reliability of the model is poor.

We conclude that, due to the asymptotic properties of the algorithm used to estimate the posterior probabilities, the LM model should be recommended especially when the prediction of the latent states is one of the main interests in the data analysis. The GMM leads to select a lower number of subpopulations compared with the LM model. However, this is not always a desirable property since when the data are rich, as in the applicative example, it may not be of interest to extremely compress their information. Within the LM model it is possible to detect also a reversible transition between the latent states. On the other hand, the consideration of the time dimension in the structural form made by the GMM is inadequate to explain the latter feature of the data.

The results proposed by the applied example may be useful when the interest is to evaluate the needs of the elderly in order to prevent fast deterioration of their health, or to investigate in more depth the reasons for improved health conditions with increasing age and therefore plan specific interventions for their health.

ACKNOWLEDGMENTS

The research has been supported by the grant " Finite mixture and latent variable models for causal inference and analysis of socio-economic data " (FIRB—Futuro in Ricerca) funded by the Italian Government (RBFR12SHVV).

» REFERENCES

» Related content