Modelling a multivariate hidden Markov process on survey data

Illustrazione del processo latente di Markov multivariato per dati collezionati in indagini campionarie

Fulvia Pennoni

Abstract We show how to handle the information which is acquired in a dynamic framework when there are multiple items in a survey collected on the same individuals at different time occasions. The response variables are commonly measured on an ordinal scale and the data may show a non-monotone missing pattern. The underlying phenomenon which is related to the interest of the survey may be modelled by a latent stochastic process. The latter having dependences according to a Markov structure is able to capture the heterogeneity of the response behaviour and to account for the measurement errors that naturally arise in the survey. The maximum likelihood estimation of the model parameters allows us to handle the missing data and to take advantage of the information provided by those individuals not showing complete responses. In a similar way, it is possible to consider a counterfactual framework in which the outcomes of interest are not directly observable even for the selected treatment. Such a flexible modelling approach is illustrated with two examples based on real data.

Key words: Expectation-Maximization algorithm, latent variables, mitochondrial DNA haplogroup, observational studies, recursive algorithm, treatment effect.

1 Introduction

In many context nowadays the data are collected in oder to show the multifacet of a phenomena of interest. The survey method is still an important tool in data collection which is employed when the units are individuals and mainly when the interest lies on understanding the characteristics of some latent features over time such as the severity of a disease over time. In particular, there are some contexts in which the

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milano e-mail: fulvia.pennoni@unimib.it

use of cohort data is of main importance to capture aspects related to salient time periods. In the following, we outline briefly that in the above context individuals share an underline structure that influence the responses and therefore a convenient way to analyze the data is by considering a multivariate latent Markov (LM) model with covariates and allowing for the missing data pattern. This short summary outlines some considerations on the innovative statistical technique of the LM model with covariates [5] as a model build for data collected on surveys. In such a context, the data analysis has to be done according to the available data structure and by inspecting the data to avoid undesirable features. The proposed model helps to consider the feature above as well as to employ a suitable Expectation-Maximization (EM) algorithm [3] in order to consider subgroups of individuals sharing common features of interest. In the following, first we introduce the model set up in a general form, we skip the estimation details which can be read from the cited references. Then, we show the model formulation focusing on two examples based on real data concerning a study on cognitive decline on elders and another study on the effect of the university degree on graduates.

2 Model specification

We define $\boldsymbol{Y}_{it}$ $(Y_{i1t},\ldots,Y_{irT})$ as the response vector provided by the i-th individual $i=1,\ldots,n$ at the t time point $t=1,\ldots,T$, $j=1,\ldots,r$. Each variable is categorical with c_j categories, when c_j is equal to two the response variables are binary. A vector of the individual characteristics $\boldsymbol{X}_{it}$ may be available at each time occasion $t=1,\ldots,T$ for each individual $i=1,\ldots,n$. We argue that due to the structure of the survey this is a typical context in which the individuals share a latent process which is Markovian of first order. Such process is denoted by $\mathbf{U}_{it}$ $(\mathbf{U}=U_1,\ldots,U_T)$ and it generates a series of independences among the responses conditional on itself. The way to model the distribution of such latent process is important in oder to proper account the pattern of responses. The basic structure of the model is illustrated by a path diagram, by which it is possible to read the conditional independences (see, Figure 1).

It can be shown that such class of models well illustrated in [4] get rise to independences holding suitable properties for inference as the joint distribution is singleton transitive [15]. Therefore, additional elements that are added to the conditioning set of every existing independence statements does not violate these independences. According to this property, it is possible to handle the information of the covariates by showing how they can affect the latent stochastic process which is governed by the initial and the transition probabilities of the Markov chain. Their parameterization may be formulated according to a multinomial logit model, or a baseline category logit model up to the intention of the data analysis.

In the presence of missing data we consider a binary indicator m_{ijt} for $i=1,\ldots,n$ and $j=1,\ldots,r$ and $t=1,\ldots,T$ to state if y_{ijt} is missing or not. The conditional independence assumption between the response variables and the missing indica-

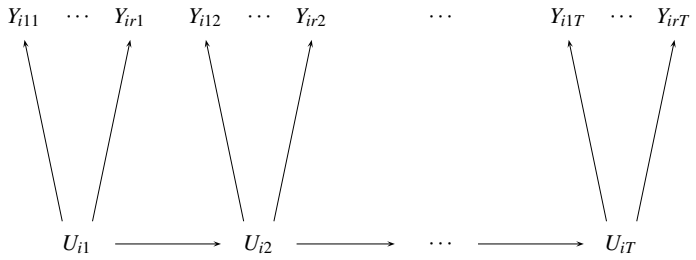

Fig. 1 Directed acyclic graph encoding the conditional independences of the multivariate latent Markov model in its basic form.

tors given the underlying latent process is still valid. Then the log-likelihood of the model may be written by considering the joint probabilities of the responses which are provided or not by the n individuals of the sample given the observed covariates. In such a context, the EM algorithm is a good numerical optimization technique to get the maximum likelihood estimates of the model parameters as it was proposed in the context of unobserved quantities [7]. Some useful recursions are employed during the iterative process which involves the maximization of the complete data log-likelihood see also [1] for more details. As the distribution of the latent process is not modelled parametrically we refer to the proposed model as a data driven model. The way the inference is conducted may be casted into the statistical techniques called empirical Bayes procedures. In fact, the selection of the subpopulations is driven by the data according to the information criteria like the Bayesian Information Criterion [13].

3 First illustrative example

The first example used to illustrate the modelling strategy is related to a cohort study developed in the United States concerning the health of 547 veterans assessed by one up to six of medical visits between 1995 and 2001[2]. Three different types of cognitive tests are considered to asses dementia: the mini mental state that examinees the global cognitive functions, the verbal fluency test that assesses the language functions (vocabulary size, naming) and the constructional praxis test concerning the visual and motor abilities. Missing responses are due to individuals that are not compliers at each visit and we relay on the assumption of ignorable missing data mechanism [12]. In Table 1 we report the distribution of the three response vari-

ables for each visit. As the cognitive impairment is considered up to a certain level of the score of the test we handle three binary response variables varying over time.

Mental test	Occasion of the interview					
	1st	2nd	3rd	4th	5th	6th
≤ 25	20.2	13.8	9.9	5.6	2.4	0.0
> 25	76.7	57.5	39.6	25.5	9.0	0.4
NA	3.1	28.7	50.5	68.9	88.7	99.6
Verbal test						
≤ 15	19.9	13.49	10.5	6.6	2.2	0.3
> 15	73.3	53.9	37.5	23.1	8.8	0.1
high	6.8	32.7	52.0	70.3	89.0	99.6
Constructional test						
≤ 3	16.1	9.6	7.2	4.9	2.7	0.0
> 3	81.9	62.7	42.4	26.5	8.4	0.4
NA	2.0	27.7	50.4	68.6	89.0	99.6

Table 1 *Percentage distribution of every response variable for the visits from 1995 to 2011.*

By considering this observational longitudinal human cohort study we aim at assessing in which way individual features, diseases and environmental exposure contributes to the cognitive resilience or decline of the elders. The research has shown that there are location places where there are no changes in mitochondrial (mtDNA) [9]. They have been called haplogroups which can also be divided according the potential different role they assume inside the mitochondria. They have also be connected with ethnicity due to the observed variability in different areas of the populated earth by human and they can be linked with migrations. MtDNA haplogroups are determined by considering the blood DNA using Taqman assays (Applied Biosystems, Foster City, CA).

To illustrate part of the available data Table 2 shows the proportion of individuals in the observed sample related to the covariates haplogroups and smoke. Table 3 illustrates the cluster composition of the haplogroups.

Covariate	Category	
haplogroup	J or T	0.15
	H or V	0.52
	K or U	0.21
	I, W or X	0.12
smoke	never	0.30
	former	0.03
	current	0.67

Table 2 *Descriptive statistics for the distributions of the covariates.*

The initial and the transition probabilities of the latent process are expresses as

Clusters of mtDNA haplopgroups	
1) J and T	Levant, Bedouin, North Eur. Eastern Eur., Indus Mediterranean
2) H and V	Europe, Western Asia, North Africa
3) K and U	West Eurasia, India sub-continent, Algeria, First Extent Middle East
4) I, X and W	Northen Eastern Europe, Amerindians, Southern Siberians, Southern Asians, Eastern Europeans, Southern East Asia

Table 3 *Classification of the observed mtDNA haplopgroups.*

$$\pi_{u|\boldsymbol{xzw}} = p(U_1 = u|\boldsymbol{X}_1 = \boldsymbol{x}, \boldsymbol{Z}_1 = \boldsymbol{z}, \boldsymbol{W}_1 = \boldsymbol{w}), \quad u = 1, \ldots, k,$$

$$\pi_{u|\bar{u}\boldsymbol{xzw}} = p(U_t = u|U_{t-1}|\boldsymbol{X}_t = \boldsymbol{x}, \boldsymbol{Z}_t = \boldsymbol{z}, \boldsymbol{W}_t = \boldsymbol{w}), \quad t = 2, \ldots, T,\ \bar{u}, u = 1, \ldots, k,$$

where $\boldsymbol{X}_t$ is the vector of the demographic characteristics of the individual which can be time-varying. In the illustrative example they are age, matrilinear ethnicity, haplogroup, years of education; $\boldsymbol{Z}_t$ is the vector of time-varying environmental factors. In the application the black carbon exposure over time is available daily. The black carbon concentration (ug/m^3) for the Massachusetts area is estimated according to a spatiotemporal land use (e.g. traffic density) regression model [8] from 83 monitoring sites. $\boldsymbol{W}_t$ is the vector of risk factors such as smoke status, hypertension, BMI (kg/m^2), diabetes.

We use a multinomial logit parameterization to account for the above covariates on the latent model. The following which is related to the initial probabilities

$$\log \frac{\pi_{u|\boldsymbol{xzw}}}{\pi_{1|\boldsymbol{xzw}}} = \alpha_{0\boldsymbol{u}} + \boldsymbol{x}'\boldsymbol{\beta}_{1u} + \boldsymbol{z}'\boldsymbol{\nu}_{1u} + \boldsymbol{w}'\boldsymbol{\tau}_{1u}, \quad u = 2, \ldots, k, \tag{1}$$

where α_{0u} is the intercept and $\boldsymbol{\beta}_{1u}, \boldsymbol{\nu}_{1u}, \boldsymbol{\tau}_{1u}$ are parameter vectors to be estimated. In a similar way as in (1) we parameterize the transition probabilities of the hidden Markov chain.

Given a sample of n independent individuals, their response vectors are $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n$ and the corresponding observed vectors of the covariates are denoted by $\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{w}}_i$. The LM model log-likelihood assumes the following expression

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\tilde{\boldsymbol{y}}_i \boldsymbol{m}_i | \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{w}}_i),$$

which involves the joint conditional probability of the observed and missing responses $\boldsymbol{m}_i$ of each individual $i, i = 1, \ldots, n$ given the observed covariates.

The LM model with $k = 2$ latent states has a log-likelihood equal to -1871.358 and a BIC index equal to 4234.464 with 78 parameters. Therefore, we identify two

main latent subpopulations of individuals having different probabilities of cognitive impairment [14]. According to the estimated conditional probabilities individuals in the first latent state have higher probability to have cognitive impairment in each of the three tests with respect to those in the second latent state. Table 4 shows the estimated intercept referred to the multinomial logit model for the initial probability. It is positive (0.886) showing that there is a general tendency towards a good cognitive status at the first visit. The log-odds referred to the four haplogroups are positive except those referred to cluster 3, indicating that those with haplogroup K and U show a worst cognitive status compared to the others at the initial visit. The estimated log-odds of the second logit referred to the smoke behaviour is higher for those who are currently smoking compared to those who never smoke, indicating that smoker show a lower cognitive status respect to nonsmokers.

		Latent state (u)
		2
intercept ($\hat{\alpha}_{0u}$)		0.824
haplogroup ($\hat{\beta}_{1u}$)	J or T	0.317
	H or V	0.220
	K or U	-0.321
	I, W or X	0.670
smoke ($\hat{\tau}_{1u}$)	never	0.820
	former	-0.598
	current	0.675

Table 4 *Estimates of the logit regression parameters affecting the initial probabilities of the latent process under the LM model with $k = 2$ latent states.*

We can select some individual of interest to evaluate their probability to belong to the best or to worst latent state at the first visit and their probability to change the cognitive health status during time by considering the estimated initial and transition probabilities. For example, if we consider the elders with haplogroup 4 which are current smokers and with an high level of education (group A) their initial probabilities are $\hat{\pi}_1^A = 0.145$ and $\hat{\pi}_2^A = 0.858$ showing that 14% of them is in the status indicating worse cognitive impairment at the first visit. If we select the corresponding group of those with the same features of group A but with a lowest level of education their initial probabilities are $\hat{\pi}_1^B = 0.174$ and $\hat{\pi}_2^B = 0.826$ showing that 17% of them is in the status indicating worse cognitive impairment. The estimated transition matrices for group A and B are shown in Table 5. From this table we can

	$\hat{\pi}^A_{u\|\bar{u}\boldsymbol{xzw}}$	
$\bar{u}$	$u=1$	$u=2$
1	0.785	0.215
2	0.084	0.916

	$\hat{\pi}^B_{u\|\bar{u}\boldsymbol{xzw}}$	
$\bar{h}$	$u=1$	$u=2$
1	0.867	0.133
2	0.092	0.908

Table 5 *Estimates of the transition probabilities for two selected groups (A, B) of individuals under the LM model with $k = 2$ latent states.*

see that there is high persistence in the same latent state for both groups but the percentage of those with a tendency to increase their cognitive decline over time is higher (22% vs 13%) for those belonging to group A which ceteris paribus are more educated. In a similar way, it is possible to evaluate the parameter estimates measuring the influence of each covariate on the transition from the first to second state.

4 Second illustrative example

Another illustrative example is related to three ordinal response variables concerning the job career of a cohort of 1,144 graduates in Lombardy. Their incomes, skills and the type of contract by which they are hired are recorded for four quarters from 2007, during the first period of the Italian economic crisis on the job market of the Lombardy region. We show that the model illustrated in Section 2 may be handled to assess causal statements in a potential outcome framework [11] in oder to establish the average treatment effect (ATE). The latter concern the effect of the degree type on the professional growth of the graduates. We consider the influence of the pre-treatment student background covariates and we implement a new way to handle the propensity score weights on the likelihood of the causal LM model. We merged three administrative archives to get the pre-treatment covariates and outcomes on graduates of four big universities of Lombardy. The treatment programs are the following: technical, architecture, economics and humanities degrees. The relative effectiveness of each degree is assessed according to the information on the mentioned features of the acquired jobs which jointly can contribute to make improvement of the human capital of each individual. The outcomes have the categories whose increasing level denotes a better job position. The pre-treatment measures assess the demographic characteristics, the high school type and the final score at the high school.

The steps we consider to estimate the ATE of the multiple treatments are the following: *i*) the pre-treatment covariates are selected according to their dependence with the treatment which is assessed by considering an ANOVA model or a chi-square test for continuous or categorical variables respectively; *ii*) the balance among treatments is established by using weights obtained by estimating a multinomial logit model; *iii*) a suitably parametrized causal LM model is estimated by maximizing thorough the EM algorithm the weighted log-likelihood; *iv*) the model selection is performed by considering the Bayesian posterior probability of a candidate model according to the observed data.

To show some results suppose we are interested on the ATE of the scientific degree at the fourth quarter of observation. We define $U_{i4}^{(z=s)}$ as the potential latent variable of individual i ($i = 1,\ldots,n$) if he/she had taken the scientific treatment. Once we have estimated the multinomial logit model for the probability of receiving the treatment given the pre-treatment covariates we notice that the unweighted mean of the final high score diploma for those attending a scientific degree is equal

to 82.183 and the weighted mean is equal to 80.46. The unweighted proportion of lyceum for those attending the scientific degree is 0.901 and the weighted one is equal to 0.795. The proposed model framework under the consistency rule and strong ignorability assumption [10] allow us to estimate the logits of the initial and transition probabilities of the hidden process and to evaluate the states according to the estimated conditional probabilities of the responses given the latent variable.

According to the BIC index we select a causal LM with 4 latent states denoting four main different subpopulations of individuals [6]. Then the conditional probabilities of the responses given the latent state 3 can be written as

$$\phi_{jy|3} = p(Y_{ijt} = y|U_{it}^{(z=s)} = 3),$$

where $j = 1,\ldots,3$, $y = 0,\ldots,2$, $i = 1,\ldots,n$, $t = 1,\ldots,T$. If we consider the estimated probabilities referred to the selected causal LM model the values are reported in Figure 2 for the latent states 3 and 4. From Figure 2 it is clear that for the professional growth the latent state 4 has to be judged better then latent state 3 as it has the highest probability of more stable contracts and of relative high earnings in a quarter. If we wish to consider the estimate of the ATE effect of the scientific treatment

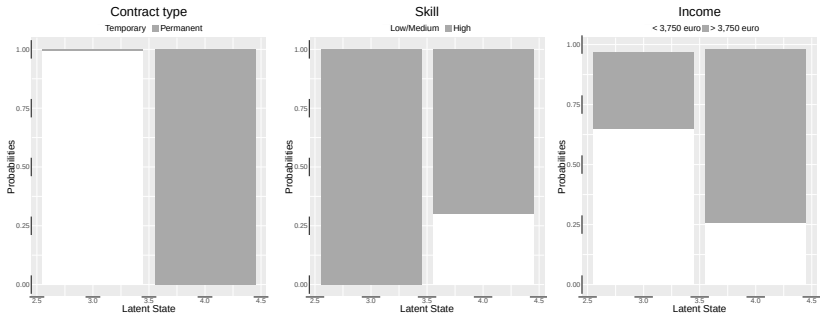

Fig. 2 *Estimated conditional probabilities ($\phi_{jy|h}$) when $h = 3,4$ of the causal LM model with $k = 4$ latent states referred to: temporary or permanent contract (Figure 1), low-medium or high skill (Figure 2) and quarterly earnings less or higher than 3750 euro (Figure 3).*

on the transition probabilities from latent state 3 to 4, we relay on the following logit parameterization

$$\log \frac{p(U_{it}^{(z=s)} = 4|U_{i,t-1}^{(z=s)} = 3)}{p(U_{it}^{(z=s)} = 1|U_{i,t-1}^{(z=s)} = 3)} = \gamma_{34} + I(z = s)\delta_4, \quad i = 1,\ldots,n,\ t = 2,\ldots,T.$$

where γ_{34} is the intercept and δ_4 is the ATE of the scientific degree with respect to technical degree (reference treatment) on the transition probabilities from $t-1$ to t (e.g. from quarter one to quarter two) to switch from latent state 3 to latent state 4. The estimated effect on the causal LM model with $k = 4$ after weighting and con-

trolling for unbalance pre-treatment covariates is $\hat{\delta}_4 = -1.507$ which is significant at 1% indicating that the scientific degree does not permit to reach the best latent state as the technical degree on the whole period of observation. The other results suggest that we can distinguish among four levels of human capital: a low level, an intermediate level with high tendency of temporary jobs with high skill level; an intermediate level with high tendency towards permanent and less skilled jobs and an high level with tendency towards permanent, high skill jobs and quite high earnings. At the beginning of the period of observation, there is a statistical significant difference of technical and economic degrees in terms of their effect on the professional growth with respect to architecture and humanities. Later on time, there is a strong significant difference between a technical degree and all the other types of degree.

It is worth mentioning that, as the model is selected on the basis of the BIC index and the maximized likelihood of the model is weighted with weights related to the propensity score, the model estimation procedure resembles the Bayesian estimation approach in which a prior is considered to locate the mode of the posterior distribution. The proposed causal formulation of the LM model can be applied to other contexts of interest such as for the secondary observational data analysis of medical treatments. In this situation, the interest may lies for example on understanding the effect of different drugs on outcomes by taking into account demographic features and previous diseases of the patients.

Acknowledgements We acknowledge the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project "*Mixture and latent variable models for causal inference and analysis of socio-economic data").*

References

1. Bacci, S., Pandolfi, S., Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, **8**, 125–145.
2. Bell, B., Rose CL., Damon, A. (1966). The veterans administration longitudinal study of healthy aging. *Gerontologist*, **6**, 179–184.
3. Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.
4. Bartolucci F., Farcomeni A., Pennoni F. (2013). *Latent Markov models for longitudinal data*. Chapman and Hall/CRC press, Boca Raton.
5. Bartolucci, F., Farcomeni, A., Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates (with discussion), *Test*, **23**, 433–465.
6. Bartolucci F., Pennoni F., Vittadini, G. (2016). Causal latent Markov models for the comparisons of multiple treatments in observational longitudinal studies. *Journal of the Educational and Behavioral statistics*, **41**, 146–179.
7. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

8. Gryparis, A., Coul, B. A., Schwartz, J., Suh, H. H. (2007). Semiparametric latent regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society, Series C*, **56**, 183-209.
9. Torroni, A., Achilli, A., Macaulay, V., Richards, M., Bandelt, H. J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*, **22**, 339-345.
10. Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
11. Rubin, D.B. (2004). Causal inference using potential outcomes: design, modeling, decisions, *The journal of the American Statistical Association*, **469**, 322–331.
12. Rubin, D.B. (1976). Inference with missing data. *Biometrika, (with discussion)*, **63**, 581–592.
13. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
14. Pennoni, F., Bartolucci, F., Baccarrelli, A. Colicio, E., Vittadini, G. (2015). Exploring the dependencies between epigenetic pathways and air pollution with the use of the latent Markov model. *4-th International Conference and Exhibition on Biometrics & Biostatistics*, 16-18 November, San Antonio, USA, **4**, p. 38.
15. Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N. and Zwiernik, P. (2015). Total positivity in Markov structures. *ArXiv:1510.0129v1*, 1–26.