Latent drop-out hidden Markov model with mixed effects

Maria Francesca Marino and Marco Alfó

Abstract We propose a class of models for the analysis of longitudinal data subject to non-ignorable drop-out. A mixed hidden Markov model for the longitudinal process is introduced with a latent drop-out class describing the influence of missingness on the response variable. A conditional generalized linear model is specified for the longitudinal profile to express dependence between observations from the same individual due to time-constant and time-varying latent characteristics. Furthermore, a latent drop-out variable is considered to explain differences between individuals having different drop-out patterns. The probability of being in one of the drop-out class is modelled through an ordinal logit model, including the time to drop-out as covariate. Parameter estimates are obtained via an EM algorithm to take into account of the presence of several (discrete and continuous) latent variables.

Key words: Hidden Markov, random effects, latent drop-out.

1 Introduction

Longitudinal data represent repeated measures from a number of sample units recorded over the time. The dependence between observations from the same individual must be taken into account to obtain valid inferences: in a regression context, this association can be ascribed to unobserved characteristics, i.e. omitted covariates whose effects can be time-constant or time-varying. In such a situations, a suitable approach is given by the so called mixed hidden Markov model, obtained by com-

Maria Francesca Marino
Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, e-mail: mariafrancesca.marino@uniroma1.it

Marco Alfó
Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, e-mail: marco.alfo@uniroma1.it

bining the features of hidden Markov and mixed effect models. See among others [2] and, for a thorough review, [11]. A common problem with longitudinal data is missingness in the form of a potentially non-ignorable drop-out; see [9] for a detailed definition of drop-out characteristics. To account for dependence between measurements and missingness, [16] propose, in a linear mixed model, a latent drop-out class variable defining a number of drop-out patterns that are meant to explain differences between subjects due to the presence of missing information. The model we propose is obtained considering a mixed hidden Markov model conditional on a latent drop-out class. This approach leads to a very flexible way to deal with longitudinal data subject to monotone missingness: the mixed hidden Markov component allow us to model dependence between observations due time-constant and time-varying latent characteristics, while the drop-out latent class helps us model existing relations between the longitudinal and the drop-out process, from a pattern mixture prospective. The plan of the paper follows. In section 2 we introduce the standard mixed hidden Markov model. Sections 3 and 4 describe the proposed model and the EM algorithm for parameters estimation. Last section contains concluding remarks and outline future developments.

2 Mixed effect hidden Markov models

Mixed effect hidden Markov models (mHMMs) are obtained combining features of hidden Markov and mixed effect models. The former assume the existence of two related processes: a latent process with a markovian structure and an observed measurement process used to describe the distribution of the response variable, conditional on the hidden state. For general references, see [4, 17]. On the other hand, mixed effect models, proposed by [7], are obtained introducing in the longitudinal response model one or more individual-specific random coefficients to capture latent individual-specific characteristics. Let $Y_i(t_{ij})$ denote the longitudinal response recorded on $i = 1, ..., r$ individuals at times $t_{ij}, j = 1, ..., n_i$ and define a homogeneous, hidden Markov chain $\{S_i(t_{ij})\}$ taking values in the finite set $\mathscr{S} = \{1, ..., m\}$. We assume that all the individuals share the same initial probability vector $\delta = (\delta_1, ..., \delta_m)$ and the same transition probability matrix $\mathbf{Q} = \{q_{hk}\}$. More specifically, δ_h represents the prior probability of starting in state h, while q_{hk} represents the probability of observing a transition from state h at time t_{ij-1} to k at time t_{ij}, with $h, k = 1, ..., m, j = 1, ..., n_i$. Finally, let us define a vector of subject-specific random parameters $\mathbf{b}_i$, following a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{D}$. As it is standard in longitudinal data analysis, we assume that the random coefficients vector $\mathbf{b}_i$ is independent from the hidden process $\{S_i(t_{ij})\}$. Mixed HMMs are based on the following main assumptions. First, the distribution of the observed responses at a given time point is influenced only by the hidden state occupied at the same time and the time-constant, individual-specific random effects. Second, conditional on the hidden state $s_i(t_{ij})$ and the random vector $\mathbf{b}_i$, observations from the same subject are conditionally independent. Based on these hypothesis, the following expression holds

$$f_y\big(y_i(t_{ij}) \mid y_i(t_{i1:j-1}), s_i(t_{i1:j}), \mathbf{b}_i; \psi\big) = f_y\big(y_i(t_{ij}) \mid s_i(t_{ij}), \mathbf{b}_i; \psi\big). \qquad (1)$$

Parameter estimation in the mixed hidden Markov models is typically done using a maximum likelihood approach. Le us denote by Φ the set of all model parameters, i.e. the initial and transition probabilities of the hidden Markov process, the longitudinal model parameters and the covariance matrix of the random effects. The observed likelihood is obtained as

$$L(\Phi) = \prod_{i=1}^{r} f(\mathbf{y}_i) = \prod_{i=1}^{r} \int \sum_{\mathbf{s}_i} f_y(\mathbf{y}_i \mid \mathbf{s}_i, \mathbf{b}_i; \psi) f_s(\mathbf{s}_i; \delta, \mathbf{Q}) f_b(\mathbf{b}_i; \mathbf{D}) d\mathbf{b}_i. \qquad (2)$$

In many cases, this expression can not be directly maximized because of the presence of latent variables and of summation over m^{n_i} terms for each unit; an EM algorithm can, instead, be used. Here, both the hidden Markov and the random effects parameters are treated as missing data and, because of the presence of continuous latent variables, numerical approximation is required. A detailed discussion of the EM algorithm is given in section 4.

3 Mixed HMMs with latent drop-out class

Suppose a longitudinal study is designed to collect repeated measures of a response variable Y on r subjects. Let the measurement process be affected by monotone missingness, i.e. some subjects leave the study before its completion time, presenting incomplete data: for each unit $i = 1, ..., r$, measurements are only available at time points $t_{ij}, j = 1, ..., n_i$. The variable t_{in_i} thus represents le length of time the $i-th$ subject participated in the study. We assume that the drop-out process may be related to the measurement process because of the presence of unobserved factors which may lead to non-ignorable missingness. In this context, one modelling approach is to assume that individuals with same drop-out time share some common unobserved heterogeneity and, according with pattern mixture models proposed by [8], the distribution of the complete responses is expressed as a mixture over different drop-out patterns. As pointed out by [16], this model specification can lead to a huge number of strata and to biased estimates: a restricted number of drop-out classes can be, instead, more suitable. To this purpose, let us introduce a drop-out latent variable $\eta_i = (\eta_{i1}, ..., \eta_{iG})$, such that $\eta_{ig} = 1, g = 1, ..., G$, if i belongs to the $g-th$ class and zero else. We suppose that units within the same drop-out class have some common latent characteristics which explain differences, in term of the response variable, due to the drop-out process. We assume that $\eta_i, i = 1, ..., r$, are multinomial random variables with ordered categories representing different drop-out propensities. The probability of having a drop-out propensity lower than a given level is modelled as a monotone function of the length of observation t_{in_i}, i.e. increases in the drop-out time progressively increase or decrease the probability of being in one of the first q latent classes. Based on these assumptions, the following ordinal logit model is specified:

$$\Pr\left(\sum_{l=1}^{q} \eta_{il} = 1 \mid t_{in_i}\right) = \frac{\exp\{\lambda_0 + \lambda_{1q} t_{in_i}\}}{1 + \exp\{\lambda_0 + \lambda_{1q} t_{in_i}\}} \quad (3)$$

where the constraints $\lambda_{11} \leq \lambda_{12}, ..., \leq \lambda_{1G-1}$ on the slope parameters hold to ensure that the probability of extreme categories increases with the width of the observation window. Let's now consider the hidden Markov process and the random coefficients vector $\mathbf{b}_i$ defined in seection 2. Based on the markovian property and considering the presence of missing data, the individual-specific hidden Markov distribution can be expressed as follows:

$$f_s(\mathbf{s}_i; \delta, Q) = \delta_{s_i(t_{i1})} \prod_{j=2}^{n_i} q_{s_i(t_{ij-1}) s_i(t_{ij})} \quad (4)$$

Given the current hidden state, the drop-out class and the random effects $\mathbf{b}_i$, observations from the same unit are independent with joint distribution

$$f_y(\mathbf{y}_i \mid \eta_i, \mathbf{S}_i, \mathbf{b}_i) = \prod_{j=1}^{n_i} f_y\big(y_i(t_{ij}) \mid \eta_i, s_i(t_{ij}), \mathbf{b}_i\big). \quad (5)$$

In particular, we assume that responses have conditional distribution in the exponential family

$$\big[Y_i(t_{ij}) \mid \eta_{ig} = 1, S_i(t_{ij}) = h, \mathbf{b}_i\big] \sim EF(\theta_{igh}(t_{ij})), \quad (6)$$

with canonical parameter described by the following regression model:

$$g[\theta_{igh}(t_{ij})] = \mathbf{x}_i(t_{ij})'\beta_h + \mathbf{z}_i(t_{ij})'\mathbf{b}_i + \mathbf{w}_i(t_{ij})'\gamma_g \quad (7)$$

Here, $\mathbf{x}_i(t_{ij})$ is a vector of covariates whose effects on the mean response are hidden state-specific. Covariates associated to class and individual-specific parameters are, instead, denoted by $\mathbf{w}_i(t_{ij})$ and $\mathbf{z}_i(t_{ij})$. Obviously, these may overlap, at least partially. Through model (7) we are able to take into account various sources of heterogeneity and dependence within subjects. First, state specific parameters β_h can be used to describe the effects of fixed covariates on the response variable due to unit-specific dynamics. In this way, we account for correlation between measures from the same unit, as in Heckman's [1] *true contagion*, see [6] for references. Second, subject and drop-out class specific parameters account for unobserved time-constant heterogeneity, i.e. sources that are specific to each sample unit or shared by members of the same drop-out class, mimicking a kind of *spurious* and *induced contagion*. Based on such modelling assumptions and indicating with ψ the vector

[1] The incidental parameters problem and the problem of initial conditions in estimating discrete time-discrete data stochastic processes and some Monte Carlo evidence. In Manski, C.F., McFadden, D.: *Structural analysis of discrete data*. (1981)

of longitudinal model parameters, the individual joint distribution of the observed measurements, given the length of observation can be factorized as

$$f(\mathbf{y}_i^o \mid t_{in_i}) = \int \sum_{\eta_i \mathbf{s}_i} f_y(\mathbf{y}_i^o \mid \eta_i, \mathbf{s}_i, \mathbf{b}_i; \psi) f_\eta(\eta_i; \lambda) f_s(\mathbf{s}_i; \delta, \mathbf{Q}) f_b(\mathbf{b}_i; \mathbf{D}) d\mathbf{b}_i. \quad (8)$$

where, $\mathbf{y}_i^o$ indicates observed response vector. As it is clear from (8), the drop-out class summarizes all the information on the dependence between the longitudinal and the missingness process. After conditioning on η, longitudinal measurements do not depend on the drop-out process which can be considered ignorable. Therefore, reliable inferences may be derived considering the observed data only.

4 Maximum likelihood estimation

Let $\Phi = (\psi, \lambda, \delta, \mathbf{Q}, \mathbf{D})$ represent the full set of model parameters; estimation may be performed by using a maximum likelihood approach. Considering modelling assumptions above, the observed log-likelihood is obtained as the sum of log individual contributions (8) and is given by:

$$\ell(\Phi) = \sum_{i=1}^{r} \log \int \sum_{\eta_i, \mathbf{s}_i} f_y(\mathbf{y}_i^o \mid \eta_i, \mathbf{s}_i, \mathbf{b}_i; \psi) f_\eta(\eta_i; \lambda) f_s(\mathbf{s}_i; \delta, \mathbf{Q}) f_b(\mathbf{b}_i; \mathbf{D}) d\mathbf{b}_i \quad (9)$$

Due to the presence of latent variables, the EM algorithm introduced by [5] is a natural choice for parameter estimation. It is based on the definition of a complete data log-likelihood where hidden states, random effects and latent drop-out class are treated as missing data. Let $u_{ih}(t_{ij})$ and $u_{ihk}(t_{ij})$ be indicator variables for the hidden Markov states. The former equals one if subject i is in state h at time t_{ij}; the latter equals one if i is in state h at time t_{ij-1} and in state k at time t_{ij}. Likewise, v_{ig} is a binary component indicator equal to one if the $i-th$ unit belongs to the $g-th$ drop-out class. The complete data log-likelihood can be written as

$$\ell_c(\Phi) = \sum_{i=1}^{r} \left\{ \sum_{h=1}^{m} u_{ih}(t_{i1}) \log \delta_h + \sum_{j=2}^{n_i} \sum_{h,k=1}^{m} u_{ihk}(t_{ij}) \log q_{hk} + \sum_{g=1}^{G} v_{ig} \log \pi_{ig} \right.$$
$$\left. + \sum_{j=1}^{n_i} \sum_{h=1}^{m} \sum_{g=1}^{G} u_{ih}(t_{ij}) v_{ig} \log f_y(y_i(t_{ij}) \mid g, h, \mathbf{b}_i; \psi) + \log f_b(\mathbf{b}_i; \mathbf{D}) \right\} \quad (10)$$

where $f_y(y_i(t_{ij}) \mid g, h, \mathbf{b}_i; \psi)$ is a shorthand for $f_y(y_i(t_{ij}) \mid \eta_{ig} = 1, S_{it} = h, \mathbf{b}_i; \psi)$ and π_{ig} indicates the probability that the $i-th$ subject belongs to the $g-th$ drop-out class. To simplify the estimation problem, let us introduce the forward and backward variables that are typically used in hidden Markov models, as proposed by [3]. Forward variables are defined as the joint probability of the longitudinal measures up to time t_{ij} for a generic individual ending up in state h, given the random effects $\mathbf{b}_i$ and the drop-out class g:

$$a_{ij}(g,h,\mathbf{b}_i) = f\left(y_i(t_{i1:j}), S_i(t_{ij}) = h \mid \eta_{ig} = 1, \mathbf{b}_i\right)$$

Backward variables are defined accordingly and represent the probability of the longitudinal sequence from t_{ij+1} to the last available observation, conditional on being in the $h-th$ state at t_{ij}, on the random effects $\mathbf{b}_i$ and the $g-th$ drop-out class:

$$b_{ij}(g,h,\mathbf{b}_i) = f\left(y_i(t_{ij+1:n_i}) \mid S_i(t_{ij}) = h, \eta_{ig} = 1, \mathbf{b}_i\right)$$

At the $r-th$ iteration of the algorithm, the E-step consists in calculating the expectation of (10) given the observed data and the current parameter estimates. As a result, we obtain

$$\begin{aligned} Q(\cdot) = \sum_{i=1}^{r} \Bigg\{ &\sum_{h=1}^{m} \hat{u}_{ih}(t_{i1}) \log \delta_h + \sum_{j=2}^{n_i} \sum_{h,k=1}^{m} \hat{u}_{ihk}(t_{ij}) \log q_{hk} + \sum_{g=1}^{G} \hat{v}_{ig} \log \pi_{ig} \\ &+ \sum_{j=1}^{n_i} \sum_{h=1}^{m} \sum_{g=1}^{G} \hat{u}_{ih}(t_{ij}) \hat{v}_{ig} \int \log f_y\left(y_i(t_{ij}) \mid g,h,\mathbf{b}_i;\psi\right) f_b\left(\mathbf{b}_i \mid \mathbf{y}_i^o, t_{in_i}, \Phi^{(r)}\right) d\mathbf{b}_i \\ &+ \int \log f_b(\mathbf{b}_i \mid \mathbf{D}) f_b\left(\mathbf{b}_i \mid \mathbf{y}_i^o, t_{in_i}, \Phi^{(r)}\right) d\mathbf{b}_i \Bigg\} \end{aligned} \tag{11}$$

where $\hat{u}_{ih}(t_{ij}), \hat{u}_{ihk}(t_{ij})$ and $\hat{v}_{ig}$ are the posterior probabilities of the indicator variables defined above. By doing a little algebra, we obtain

$$\hat{u}_{ih}(t_{ij}) = \frac{\int \sum_g \pi_{ig}\, a_{ij}(g,h,\mathbf{b}_i)\, b_{ij}(g,h,\mathbf{b}_i)\, f_b(\mathbf{b}_i;\mathbf{D}) d\mathbf{b}_i}{\int \sum_{hg} \pi_{ig}\, a_{ij}(g,h,\mathbf{b}_i)\, b_{ij}(g,h,\mathbf{b}_i)\, f_b(\mathbf{b}_i;\mathbf{D}) d\mathbf{b}_i}$$

$$\hat{u}_{ihk}(t_{ij}) = \frac{\int \sum_g \pi_{ig}\, a_{ij-1}(g,h,\mathbf{b}_i)\, q_{hk}\, f_y\left(y_i(t_{ij}) \mid g,k,\mathbf{b}_i\right)\, b_{ij}(g,k,\mathbf{b}_i)\, f_b(\mathbf{b}_i;\mathbf{D}) d\mathbf{b}_i}{\int \sum_{hkg} \pi_{ig}\, a_{ij-1}(g,h,\mathbf{b}_i)\, q_{hk}\, f_y\left(y_i(t_{ij}) \mid g,k,\mathbf{b}\right)\, b_{ij}(g,k,\mathbf{b}_i)\, f_b(\mathbf{b}_i;\mathbf{D}) d\mathbf{b}_i}$$

$$\hat{v}_{ig} = \frac{\int \sum_h \pi_{ig}\, a_{in_i}(g,h,\mathbf{b}_i)\, f_b(\mathbf{b}_i;\mathbf{D}) d\mathbf{b}_i}{\int \sum_{hg} \pi_{ig} a_{in_i}(g,h,\mathbf{b}_i)\, f_y(\mathbf{b}_i;\mathbf{D}) d\mathbf{b}_i}$$

As it is clear, computation of the quantities above, as well as the expected values with respect to the posterior distribution of the individual random coefficients in (11), requires the calculation of multiple integrals which can not be solved analytically, but should be numerically approached. For a detailed description of numerical approximations see [14]. For this purpose, we employ an adaptive Gaussian quadrature (AGQ) rule: each integral is approximated through a weighted summation over a pre-specified number of quadrature points. At each step of the algorithm, these locations are centered and scaled with respect to the posterior modes and curvatures of the random effects distribution. See for references [10, 13]. If compared to the standard Gauss-Hermite approach, AGQ improves the goodness of the approximation and reduces the number of iterations required; if compared to the Monte Carlo EM used by [2] in the mixed hidden Markov model contex, it results computationally more efficient, requiring a lower computing time. As pointed out by

[1], an important issue with Monte Carlo EM is the number of points to sample to approximate adequately the integrals: using too few nodes can lead to a poor approximation due to Monte Carlo error, while too many nodes can significantly increase the computational time required to obtain the approximation. Alternatively, to decrease the computational effort, a pseudo-adaptive quadrature rule can be adopted, as suggested by [15]. The M-step of the algorithm consists in maximizing (11) with respect to model parameters; due to independence assumptions and parameter separability, the maximization problem can be partitioned into different sub-problems, i.e. the maximization with respect to the hidden Markov, the longitudinal and the latent drop-out class parameters, as well as the covariance matrix of the random effects. Regarding the Markov chain parameters, initial and transition probabilities are estimated by

$$\hat{\delta}_h = \frac{\sum_{i=1}^{r} \hat{u}_{ih}(t_{i1})}{r}, \qquad \hat{q}_{hk} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} \hat{u}_{ihk}(t_{ij})}{\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{k=1}^{m} \hat{u}_{ihk}(t_{ij})}. \tag{12}$$

Estimation of the longitudinal and the latent drop-out class parameters reduces to traditional maximization problems in generalized linear (mixed) models, i.e. finding the zeros of a weighted score function, with weights related to posterior distribution of some suitable latent variable. For the longitudinal response, the weights are obtained as a product of the posterior probabilities of the hidden states, the latent drop-out class and the random effects vector, while, for the drop-out class model, the weights are given just by the posterior probabilities of the latent drop-out classes $\hat{v}_{ig}$. Finally the covariance matrix $\mathbf{D}$ for the random coefficients is estimated via restricted maximum likelihood as

$$\hat{\mathbf{D}} = n^{-1} \sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i'. \tag{13}$$

The E and the M-steps of the algorithm are iterated until convergence; the observed information matrix is computed to obtain standard errors of the parameter estimates. As frequently happens in presence of latent variables, the EM algorithm can lead to local maxima. One way to ease the problem is to run it using multiple random starting points and then choose the best model with respect to likelihood values.

5 Conclusion and further developments

We propose a latent drop-out hidden Markov model with Gaussian random effects to describe the dynamics of a longitudinal response subject to informative missingness. If compared with existing literature, our model allows to simultaneously describe the dependence between the measurement and the drop-out process, as well as association between observations due to subject-specific omitted covariates with or without a time structure. We adopt an EM approach to obtain parameter estimates and adaptive Gaussian quadrature rules to deal with the presence of multidimensional integrals, in contrast with the Monte Carlo EM algorithm used in [2] and the finite mixture approach of [12]. In our development, we have assumed orthogonality between all the latent variables introduced in the model; obviously, this assumption could be relaxed by allowing some kind of association between variables to improve model flexibility. A potential influence of the missing process on the hidden Markov chain can be considered and represent a new interesting development. A regression approach can also be adopted to model the transition probabilities, allowing to take into account intra-subjects heterogeneity.

References

1. Hartzel, J., A. Agresti, e B. Caffo: Multinomial Logit Random Effects Models. Statistical Modelling, **1**: 81–102. (2001)
2. Altman, R.J.: Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. J. Amer. Statist. Assoc. 102, **477**, 201–210. (2007)
3. Baum, L.E.; Petrie, T.; Soules, G.; Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. **41** 164–171. (1970)
4. Cappé, O.; Moulines, E.; Rydén, T. *Inference in hidden Markov models*, Springer: New York, 2005.
5. Dempster, A. P.; Laird, N. M.; Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. With discussion. J. Roy. Statist. Soc. Ser. B 39, **1**, 1–38. (1977)
6. Manski, C.F., McFadden, D.: *Structural analysis of discrete data*. Cambridge, Mass: MIT Press, 1981.
7. Laird, N.M.; Ware, J.H.: Random effects models for longitudinal data. Biometrics, **38**, 963–974. (1982)
8. Little, R.: Patten-Mixture models for multivariate incomplete data. J. Amer. Statist. Assoc. **88**, 125-134. (1993)
9. Little, R., Rubin, D. *Statistical analysis with missing data,2nd edition*, Wiley, New York, 2001.
10. Liu, Q.; Pierce, D.A.: A note on Gauss-Hermite quadrature. Biometrika 81, **3**, 624–629. (1994)
11. Maruotti, A.: Mixed hidden Markov models for longitudinal data: an overview. International Statistical Review, **79**, 427-454. (2011)
12. Maruotti, A., Rydén, T.: A semiparametric approach to hidden Markov models under longitudinal observations. Statist. Comput., **19**, 381–393. (2009)
13. Pinheiro, J.C.; Bates, D.M.: Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. Journal of Computational and Graphical Statistics, **4**, 12-35. (1995)
14. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical recipes. The art of scientific computing. Third edition.* Cambridge University Press, Cambridge, 2007.
15. Rizopoulos, D.: Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. Comput. Statist. Data Anal. 56 , no. **3**, 491–501. (2012)
16. Roy, J.: Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. Biometrics 59, **4**, 829–836. (2003)
17. Zucchini, W.; MacDonald, I.L. *Hidden Markov models for time series. An introduction using R. Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2009.