Dynamic random coefficient based drop-out models for longitudinal responses

Modelli a coefficienti casuali dinamici per risposte longitudinali affette da drop-out non-ignorabile

M.F. Marino and M. Alfó

Abstract We propose a dynamic random coefficient based drop-out model for the analysis of longitudinal data subject to potentially non-ignorable drop-out. The presence of a non-ignorable missingess may severely bias inference on the observed data. In this framework, random coefficient based drop-out models represent an flexible approach to jointly model both longitudinal responses and missingess. We extend such an approach by allowing the random parameters in the longitudinal data process to evolve over time according to a non-homogeneous hidden Markov chain. The resulting model offers great flexibility and allows us to efficiently describe both between-outcome and within-outcome dependence.

Abstract *Gli studi longitudinali sono spesso caratterizzati dalla presenza di dati mancati dovuti ad alcuni individui che lasciano lo studio anticipatamente. Quando il meccanismo che conduce al dato mancante è non ignorabile, è possibile giungere a conclusioni inferenziali valide solo modellando congiuntamente due outcome: il processo longitudinale ed il processo generatore del dato mancate stesso. A questo scopo, si propone un modello di regressione per dati longitudinali soggetti ad a drop-out potenzialemente non ignorabile in cui coefficienti casuali tempo-constanti e tempo-variabili vengono congiuntamente presi in considerazione. Questo permette di modellare in maniera opportuna sia la dipendenza esistente tra le misurazioni di ripetute di uno stesso outcome per una stessa unit statistica, sia la dipendenza esistente tra outcome diversi.*

Key words: Hidden Markov models, nonparametric maximum likelihood, random effects, missingness,

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Universitá degli Studi di Firenze, e-mail: mariafrancesca.marino@unifi.it

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

1 Introduction

Longitudinal studies are frequently affected by drop-out. If the selection of individual staying in the study still depends on (future) unobserved responses once conditioning on the observed data, the missingness mechanism is said to be non-ignorable [9]. In this respect, to obtain valid inference, missingness should be taken in explicit account.

Different alternatives are available in the literature to deal with non-ignorable drop-outs [8]. Among them, random coefficient based drop-out model [RCBDM - 7] represent an interesting approach. They allow for the presence of two different sets of individual-specific random parameters for the longitudinal and the missing data process, respectively. These capture the dependence between repeated measurements from the same individual (within-individual dependence). The corresponding joint distribution provides instead a measure of dependence between the longitudinal and the missingness process (between-outcomes dependence).

When dealing with longitudinal data, the assumption of time-constant, individual-specific, sources of unobserved heterogeneity may be too restrictive [1]. Starting from the proposal by [10], we introduce a dynamic random coefficient based drop-out model, where time-varying random parameters are considered to model the longitudinal outcome. To explain our proposal, we assume that the dependence between the longitudinal and the missing data process is captured by an individual-specific *upper-level* mixture. Also, to describe the dependence within profiles, we consider two further sets of random parameters. For the longitudinal outcome, individual-specific, time-varying, random parameters that evolve over time according to a non-homogeneous hidden Markov chain are exploited. On the other hand, for the missing data outcome, we consider individual-specific, time-constant, random parameters identifying non-homogeneous propensities to stay into the study.

The proposed model is applied to the Leiden 85+ dataset where the effect of demographic and genetic factors on the evolution of cognitive functioning in elder people is of main interest [3]. Due to poor health conditions or death, individuals enrolled in the study may present incomplete sequences. We show how the proposed model specification may be fruitfully exploited to derive valid inference on the parameters of interest.

2 Motivating example: the Leiden 85+ study

The Leiden 85+ study is a longitudinal study conducted by the Leiden University Medical Centre in the Netherlands, with the aim at analysing the evolution of cognitive functioning in the elderly. The study entails Leiden inhabitants who turned 85 years old between September 1997 and September 1999. The sample is made by 541 elderly who were followed for six consecutive yearly visits until they reached 90 years of age. Patient conditions were assessed via the Mini Mental Status Examination [MMSE, 6] index taking values between 0 and 30 with higher values corre-

sponding to better cognitive skills. The aim of the study is that of identifying demographic and genetic factors influencing the dynamics of cognitive functioning and healthy aging. To this purpose, the following covariates were measured: age, *gender*, *educational status*, and *APOE genotype*. The latter identifies the Apolipoprotein E genotype of the patient; in particular, ε_4 allele is known to be linked to the risk of dementia. Due to the design of the study, a number participants present incomplete responses (i.e. drop-out), because of poor health conditions or death.

Preliminary analysis show that MMSE values generally reduce with time but such a trend is more evident for subjects dropping out prematurely. Such a finding poses the question on whether the process leading to missing data may be ignored. In the next section, we will introduce a dynamic RCBDM to account for both the potential dependence between the longitudinal data process and the drop-out mechanism and the within-profile dependence.

3 The dynamic RCBDM

Let us suppose a longitudinal study is designed to collect measures for a response variable $Y_{it}, i = 1,\ldots,n, t = 1,\ldots,T$, on a sample of n individuals at T time occasions and let $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iT})$ denote the vector of individual response sequences. As it it is frequent when dealing with longitudinal studies, some individuals in the sample may drop-out prematurely and, thus, may present incomplete sequences. In this framework, let $\mathbf{R}_i = (R_{i1},\ldots,R_{iT_i^*})'$ indicate the T_i^*-dimensional missing data vector, where $T_i^* = \min(T_i+1, T)$ and T_i denotes the number of available measurements for the i-individual. R_{it} is defined as a binary variable with $R_{it} = 0$ if the i-th individual drops-out from the study between occasion $t-1$ and t and $R_{it} = 1$ otherwise.

Furthermore, let $Z_{it} \in \{1,\ldots,G\}$ and $U_i \in \{1,\ldots,K\}$ be two individual-specific, discrete, latent variables influencing the longitudinal and the missing data process, respectively. As it is clear, while the latter variable is assumed to depend on the individual i only, the former variable is individual- and time-specific. This allows us to capture sources of unobserved dynamics that influence Y_{it} and that would be barely captured by a time-constant latent variable.

We assume that the longitudinal outcome Y_{it} only depends on the corresponding latent variable Z_{it} and, conditional on the vector $\mathbf{Z}_i = (Z_{i1},\ldots,Z_{iT})$, the elements of $\mathbf{Y}_i$ are independent, with joint (conditional) density given by

$$f(\mathbf{y}_i \mid \mathbf{Z}_i = \mathbf{z}_i) = \prod_{t=1}^{T} f(y_{it} \mid Z_{it} = z_{it}).$$

Similarly, we assume that conditional on the latent variable U_i, missingness indicators are independent and the corresponding joint (conditional) density is

$$f(\mathbf{r}_i \mid U_i = u_i) = \prod_{t=1}^{T_i^*} f(r_{it} \mid U_i = u_i).$$

To describe the effect of the observed covariates on the outcomes (Y_{it}, R_{it}), the following regression models are also defined:

$$\begin{cases} g[\mathrm{E}(Y_{it} \mid Z_{it} = g)] = \zeta_g + \mathbf{x}'_{it}\beta, \\ \mathrm{logit}[\Pr(R_{it} = 0 \mid U_i = k)] = \xi_k + \mathbf{w}'_{it}\gamma. \end{cases}$$

In the expressions above, $g(\cdot)$ represents an appropriate link function, while the parameters β and γ describe the effects of observed covariates, $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$, on Y_{it} and R_{it}, respectively. Also, $\zeta_g, g = 1, \ldots, G$, denotes the value of the random intercept in the longitudinal data model when $Z_{it} = g$. To simplify the interpretation of such parameters, we introduce the following ordinal constraint:

$$\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_G, \tag{1}$$

so that lower values of Z_{it} correspond to lower values for the longitudinal response. Last, $\xi_k, k = 1, \ldots, K$, denotes the discrete random intercept associated to the missing data process when $U_i = k$.

Following an approach similar to that suggested by [10], we model the dependence between Z_i and U_i and, therefore, between the longitudinal and the missing data process, by considering a discrete *upper-level* latent variable, V_i, defined on the support set $\{1, \ldots, H\}$, with $\tau_h = \Pr(V_i = h), h = 1, \ldots, H$. In particular, we assume that, conditional on $V_i = h$, the latent variables $\mathbf{Z}_i$ and U_i are independent with joint distribution described by the following (association) model:

$$f(\mathbf{Z}_i, U_i) = \sum_{h=1}^{H} \tau_h \left[\Pr(\mathbf{Z}_i = \mathbf{z}_i \mid V_i = h) \Pr(U_i = u_i \mid V_i = h)\right].$$

With the aim of accounting for time-varying sources of unobserved heterogeneity influencing the longitudinal data process, we assume that, conditional on the h-th component of the upper-level mixture, that is conditional on $V_i = h$, the latent variables Z_{it} evolve over time according to a first order hidden Markov chain with initial probability vector δ_h and transition probability matrix $\mathbf{Q}_h$, with $h = 1, \ldots, H$.

3.1 Reducing model complexity

As it can be noticed, the adopted parameterization is quite complex. This could lead to numerical difficulties when deriving the corresponding maximum likelihood estimates. In order to reduce the number of parameters, we follow an approach similar to that by [4] and specify δ_h and $\mathbf{Q}_h$ via a global logit parameterization. This choice is motivated by the constraints specified in equation (1) which, in turn, lead to considering the latent variable Z_{it} having as ordinal. In this framework, initial

probabilities of the hidden Markov chain are defined according to the model

$$\log \frac{\delta_{g|h} + \dots \delta_{G|h}}{\delta_{1|h} + \dots + \delta_{g-1|h}} = \alpha_{0g} + \psi_{0h}, \tag{2}$$

with $h = 1, \dots, H$ and $g = 2, \dots, G$. For identifiability purposes, we set $\psi_{01} = 0$, so that the number of parameters to be estimated reduces to $(G-1)+(H-1)$. On the other hand, transition probabilities are modelled according to the following ordinal logit:

$$\log \frac{q_{gg'|h} + \dots + q_{Gg'|h}}{q_{1g'|h} + \dots + q_{g-1g'|h}} = \alpha_{1gg'} + \psi_{1h}, \tag{3}$$

with $h = 1, \dots, H$, $g = 1, \dots, G$, and $g' = 2, \dots, G$. As above, to ensure parameter identifiability, we set $\psi_{11} = 0$, so that $G(G-1)+(H-1)$ parameters need to be estimated.

4 Model inference

Let θ denote the vector of all model parameters. Estimation of such parameters can be carried out via a maximum likelihood approach. Due to the local independence assumption within and between the longitudinal and the missing data responses, $\mathbf{Y}_i$ and $\mathbf{R}_i$, inference may be based on the following observed data likelihood:

$$L(\theta) = \prod_{i=1}^{n} \sum_{h}^{H} \tau_h \left\{ \sum_{Z_{i1} \cdots Z_{iT_i}} \left[\prod_{t=1}^{T_i} f(y_{it} \mid Z_{it} = z_{it}) \, \delta_{z_{i1}|h} \prod_{t=2}^{T} q_{z_{it-1} z_{it}|h}) \right] \times \right.$$
$$\left. \times \left[\prod_{t=1}^{T_i^*} \sum_{u_i} f(r_{it} \mid U_i = u_i) \, \pi_{u_i|h} \right] \right\},$$

To avoid multiple summations over all possible realisations of the hidden chain, $Z_{i1}, \dots, Z_{iT_i}$, we may rely on the EM algorithm [5].

In this framework, two separated steps need to be alternated. In the E-step, we need to compute expected value of the complete data log-likelihood, conditional on the observed data and the current value of parameter estimates. Such a computation can be consistently simplified by extending the standard forward-backward algorithm [2] which is typically used in the hidden Markov model framework. In the M-step, we need to maximize the expected value of the complete data log-likelihood with respect to model parameters. The E- and the M-steps are iterated until convergence. As it is frequent when dealing with discrete latent variables, to avoid local maxima or spurious solutions, we may consider a multi-start strategy based on both deterministic and random solutions. Also, the number of upper- and lower-level components/states is treated as fixed and known. The algorithm is run

for varying choices of (G,K,H) and the best model is chosen via standard model selection techniques.

References

[1] F. Bartolucci and A. Farcomeni. A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. *Biometrics*, 71:80–89, 2015.

[2] L. E Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[3] A. Bootsma-Van Der Wiel, E. Van Exel, A.J.M. De Craen, J. Gussekloo, A.M. Lagaay, D.L. Knook, and R.G.J. Westendorp. A high response is not essential to prevent selection bias: results from the leiden 85-plus study. *Journal of clinical epidemiology*, 55:1119–1125, 2002.

[4] R. Colombi and A. Forcina. Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, pages 1007–1019, 2001.

[5] A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.

[6] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12:189–198, 1975.

[7] Nisha C. Gottfredson, Daniel J. Bauer, and Scott A. Baldwin. Modeling change in the presence of nonrandomly missing data: Evaluating a shared parameter mixture model. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2):196–209, 2014.

[8] R. J.A. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley, 2002.

[9] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[10] Alessandra Spagnoli and Marco Alfò. Random coefficient based dropout models: a finite mixture approach. In *46th Scientific Meeting of the Italian Statistical Society*, 2012.