A non-homogeneous hidden Markov model for partially observed longitudinal responses

Maria Francesca Marino
mariafrancesca.marino@unifi.it

Marco Alfó
marcoalfo@uniroma1.it

Abstract

Dropout represents a typical issue to be addressed when dealing with longitudinal studies. If the mechanism leading to missing information is non-ignorable, inference based on the observed data only may be severely biased. A frequent strategy to obtain reliable parameter estimates is based on the use of individual-specific random coefficients that help capture sources of unobserved heterogeneity and, at the same time, define a reasonable structure of dependence between the longitudinal and the missing data process. We refer to elements in this class as random coefficient based dropout models (RCBDMs). We propose a dynamic, semi-parametric, version of the standard RCBDM to deal with discrete time to event. Time-varying random coefficients that evolve over time according to a non-homogeneous hidden Markov chain are considered to model dependence between longitudinal responses recorded from the same subject. A separate set of random coefficients is considered to model dependence between missing data indicators. Last, the joint distribution of the random coefficients in the two equations helps describe the dependence between the two processes. To ensure model flexibility and avoid unverifiable assumptions, we leave the joint distribution of the random coefficients unspecified and estimate it via nonparametric maximum likelihood. The proposal is applied to data from the Leiden 85+ study on the evolution of cognitive functioning in the elderly.

Keywords: Dropout; Finite Mixture, Latent Markov Model; Missingness; Nonparamet-
1 Introduction

Missingness represents a frequent issue to be handle in longitudinal studies, as some participants may not be available at all intended time occasions and, therefore, may present incomplete data records. Monotone missingness represents the most frequent type of non-participation, with some individuals leaving the study prematurely and having a zero probability to re-enter. Rubin (1976) introduced a well known taxonomy for missing data mechanisms which can be either based on the potential link between the longitudinal and the drop-out process or on the impact of missing data on parameter estimates in the longitudinal data model. In this respect, the drop-out may be non-ignorable in the sense that, even after conditioning on the observables (both covariates and responses), the participation to the study still depends on future (potentially unobserved) response values. Obviously, such a phenomenon may bias the study design and the resulting inference.

Different modeling approaches to deal with such non-ignorable missingness may be found in the statistical literature. Little (1995) and Little and Rubin (2002) distinguish two broad classes of models to account for missing responses: selection and-pattern mixture models. Differences rely on the factorization of the joint distribution for the longitudinal and the missing data process which, in turn, may be linked to different interpretation of the dependence between the two processes. A further modeling alternative is based on the inclusion of individual-specific, typically Gaussian, random coefficients in the model specification to capture both the dependence between repeated measurements from the same individual (within-individual dependence) and between the longitudinal and the missing data process (between-outcomes dependence). In the simplest case, the same set of random coefficients is shared by the two processes, with the resulting model being usually referred to as shared parameter model (e.g. Wu and Carroll, 1988; Wu and Bailey, 1988; De Gruttola and Tu, 1994). Missing data models
with nonparametric (discrete) random effects/coefficients were proposed by Alfò and Aitkin (2000), Roy (2003), Tsonaka et al. (2009), and Alfò and Maruotti (2009). As noticed by Bartolucci and Farcomeni (2015), a limitation of standard shared parameter models is that random coefficients are assumed to be constant over time. This may be appropriate with dealing with short time series only. For this reason, they proposed a shared parameter model for multivariate longitudinal responses and a (discrete) time to event, based on the use of time-varying and time-constant (discrete) random intercepts which are both shared by the longitudinal and the missingness model.

However, an implicit assumption of shared random parameter models is that the primary outcome and the drop-out mechanism are influenced by the same sources of unobserved heterogeneity. Clearly, this assumption may be restrictive in some cases and, above all, cannot be verified by looking at the observed data only. A more general approach to deal with non-ignorable missingness is that of considering two separate sets of random coefficients in the equations for the longitudinal and the drop-out process. The corresponding joint distribution is used to describe the dependence between these latter; see, among others, Alfò and Maruotti (2009), Creemers et al. (2010, 2011), Gottfredson et al. (2014), and Barrett et al. (2015). While a Gaussian multivariate distribution is frequently assumed also in this case, a nonparametric specification can be considered to improve model flexibility as in Spagnoli et al. (2017). As with shared parameter models, the assumption of time-constant random coefficients may be restrictive in some cases (Bartolucci and Farcomeni, 2015). While a possible strategy to describe how outcomes evolve over time may consist in including in the model specification some function of time associated to fixed or random parameters, such an approach would help us describe “well-shaped” (e.g. polynomial) dynamics only. A more flexible and appealing approach may be based on the use of a Hidden Markov Model (HMM) formulation (Zucchini and MacDonald, 2009; Bartolucci et al., 2013).

In this respect, we introduce a random coefficient based hidden Markov model for longitudinal responses subject to possible non-ignorable dropouts. To describe the dependence within profiles, that is between longitudinal responses and missingness
indicators from the same subject, we exploit two different sets of random coefficients. For the longitudinal outcome, we consider time-varying (discrete) random coefficients that evolve over time according to a non-homogeneous hidden Markov chain. These allows us to capture differential dynamics in the longitudinal responses over time. On the other hand, for the missing data indicator, we consider time-constant (discrete) random coefficients that allows us to identify differential propensities to stay into the study. Dependence between the two profiles is modelled via an upper-level latent class variable, with unspecified distribution which is estimated in a nonparametric maximum likelihood framework.

Our proposal is applied to data from the Leiden 85+ study, where the effect of demographic and genetic factors on the evolution of cognitive functioning for the elderly represents the main target of inference. Due to poor health conditions or death, a number of individuals enrolled in the study present incomplete data sequences. Inference based on the observed data only may lead to biased parameter estimates as, due to the study design, dependence between the drop-out and the unobserved responses is quite reasonable.

The paper is organized as follows. In Section 2, we describe the Leiden 85+ study. In Section 3, we briefly introduce random coefficient based drop-out model, while in Section 4 we describe the proposed model specification. Section 5 entails the EM algorithm for maximum likelihood parameter estimation and the procedure to derive the estimated standard errors. In Section 6, we describe the results from the application of the proposed approach to the Leiden 85+ data. The last section, contains concluding remarks.

2 The Leiden 85+ study

The Leiden 85+ study is a longitudinal study conducted by the Leiden University Medical Center in the Netherlands, with the aim of analysing the evolution of cognitive functioning in the elderly. The study entailed Leiden inhabitants who turned 85
years old between September 1997 and September 1999. Out of 705 subjects who were eligible to the study, 14 died before they could be enrolled, 92 refused to participate into the study, and 38 refused to provide blood sample. At the end, 561 elderly were followed up to six consecutive yearly visits until 90 years of age. See Bootsma-Van Der Wiel et al. (2002) and Van Vliet et al. (2010) for further details on the study. Participants’ cognitive conditions were assessed via the Mini Mental Status Examination index (MMSE, Folstein et al., 1975), which is obtained by evaluating the attention, the orientation, the language skills and the ability of the participant to perform simple actions. The corresponding questionnaire is based on 30 binary (sub-) items grouped into seven different cognitive areas. Interviewers assigned value 1 to correct answers, so that the MMSE scores (defined as the sum over the 30 sub-items) can take all integer values in the interval $[0,30]$. The study aims at identifying demographic and genetic factors that influence cognitive functioning and healthy ageing. To this purpose, the following covariates were measured at the study start: gender, educational status – a binary variable equal to 1 for not less than 7 years of schooling, and APOE genotype – a categorical variable identifying the Apolipoprotein E genotype of the patient. The three largest genotype groups ($\epsilon_2, \epsilon_3, \epsilon_4$) and their products were considered, leading to four different categories for this variable: $APOE_{22-23}, APOE_{24}, APOE_{33}$, and $APOE_{34-44}$. APOE genotype is known to play some role in aging. In particular, $\epsilon_4$ allele is known to be linked to an increased risk for dementia, whereas $\epsilon_2$ allele carriers are relatively protected.

Due to the design of the study, a number of participants present incomplete responses (i.e. dropout), due to poor health conditions or death. We report in Figure 1 the distribution of the MMSE scores at each visit stratified by whether subjects dropout between the current and the next occasion. Based on this figure, it is clear that the overall response reduces with time; that is, as expected, cognitive skills reduce with people getting older. However, from this graph, it is also evident that MMSE values tend to reduce faster for subjects dropping out prematurely. Furthermore, difference with those remaining longer under observation is more evident when the drop-out is
observed at the beginning of the study. Such a finding poses the question on whether the process leading to missing data can be ignored.

### 3 Random coefficient based drop-out model

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. As it is frequent with longitudinal studies, some individuals in the sample may dropout prematurely and, thus, present incomplete sequences. In this framework, let \( R_i = (R_{i1}, \ldots, R_{iT})' \) denote the \( T \)-dimensional missing data vector, with \( R_{it} = 0 \) if the \( i \)-th subject is available at time occasion \( t \) and \( R_{it} = 1 \) otherwise. As we focus on monotone missingness, \( R_{it} = 1 \Rightarrow R_{it'} = 1, t, t' = 1, \ldots, T, \) with \( t' > t \). Let \( b_i = (b_{i1}, \ldots, b_{im})' \) denote a vector of individual-specific random coefficients in the longitudinal data model, with \( E(b_i) = 0 \) and \( \text{Cov}(b_i) = \Sigma_b, \ i = 1, \ldots, n \). As it is standard when dealing with random coefficients, we assume that, conditional on \( b_i \), longitudinal measures coming from the same subject are independent. Denoting by \( Y_i = (Y_{i1}, \ldots, Y_{iT})' \) the vector of responses associated to the \( i \)-th subject in the sample, the corresponding joint
(conditional) density is given by:

\[ f_y(y_i \mid b_i) = \prod_{t=1}^{T} f_y(y_{it} \mid b_i). \]

Also, let \( c_i = (c_{i1}, \ldots, c_{ir})' \) denote a further set of individual-specific random coefficients that describe dependence in the missing data process, with \( E(c_i) = 0 \) and \( \text{Cov}(c_i) = \Sigma_c. \) The notation is completed by defining \( T_i^* = \min(T_i + 1, T) \), where \( T_i = T - \sum_{t=1}^{T} R_{it} \) denotes the number of available measures for individual \( i = 1, \ldots, n. \)

Conditional on \( c_i \), random variables \( R_{i1}, \ldots, R_{iT_i^*} \) are assumed to be independent, with joint (conditional) density given by:

\[ f_r(r_i \mid c_i) = \prod_{t=1}^{T_i^*} f_r(r_{it} \mid c_i). \]

An essential feature of RCBDMs is the assumption of conditional independence between the longitudinal and the missing data process. That is, conditional on the individual-specific vectors \( (b_i, c_i) \), \( Y_i \) and \( R_i \) are assumed to be independent. In this framework, denoting by \( Y_i^o \) and \( Y_i^m \) the observed and the missing data in the individual sequence \( Y_i = (Y_i^o, Y_i^m) \), the individual observed-data likelihood can be obtained as:

\[ L_i(\cdot) = \int_B \int_C \int_{Y_i^m} f_y(y_i^o \mid b_i) f_r(r_i \mid c_i) f_{b,c}(b_i, c_i) \\ \times dy_i^m \, db_i \, dc_i, \]

\[ = \int_B \int_C f_y(y_i^o \mid b_i) f_r(r_i \mid c_i) f_{b,c}(b_i, c_i) \, db_i \, dc_i \]

\[ = \int_B \int_C f_y(y_i^o \mid b_i) f_r(r_i \mid c_i) f_{b,c}(b_i, c_i) \, db_i \, dc_i \]

where \( f_y(y_i^o \mid b_i) = \prod_{t=1}^{T_i} f_y(y_{it} \mid b_i). \)

While it is common to assume that individual-specific random coefficients \( (b_i, c_i) \) follow a specific parametric (usually Gaussian) distribution, a more flexible alternative recasts this problem in a finite mixture framework. Let \( Z_i \) denote an individual-specific latent variable defined on the support set \( \{1, \ldots, G\} \) with masses \( \pi_g = \Pr(Z_i = g). \) In the \( g \)-th component of the finite mixture, random coefficients \( b_i \) and \( c_i \) take value \( \zeta_g \) and \( \xi_g \), respectively, \( g = 1, \ldots, G. \) In this context, the individual contribution to the
observed data likelihood in equation (1) can be re-formulated as

\[ L_i(\cdot) = \sum_{g=1}^{G} f_y(y_i^g \mid Z_i = g) f_r(r_i \mid Z_i = g) \pi_g. \] (2)

Although the finite mixture representation is based on a robust and consistent (nonparametric) estimate of \( f_{b,c}(b_i, c_i) \), it is evident that this approach is not defined to (effectively) distinguish \textit{within} and \textit{between} profile dependence. This is due to the intrinsic unidimensionality of the discrete latent variable \( Z_i \). For this reason, Spagnoli et al. (2017) proposed a generalization of the model above, where two distinct sets of latent variables, with a possibly different number of categories, are considered. A first variable, \( Z_i \in \{1, \ldots, G\} \), is used to account for the dependence between longitudinal responses from the same subject, while a second one, \( U_i \in \{1, \ldots, K\} \), allows to model the dependence in the individual sequence of missing data indicators. This formulation leads to the following expression for the individual contribution to the observed data likelihood:

\[ L_i(\cdot) = \sum_{g=1}^{G} \sum_{k=1}^{K} f(y_i^g \mid Z_i = g) f(r_i \mid U_i = k) \pi_{gk}, \]

where \( \pi_{gk} = \Pr(Z_i = g, U_i = k) \) is used to model dependence between \( Y_i \) and \( R_i \).

Obviously, when \( \pi_{gk} = \pi_g \cdot \pi_k \), with \( \pi_g = \sum_{k=1}^{K} \pi_{gk} \) and \( \pi_k = \sum_{g=1}^{G} \pi_{gk} \), independence holds and maximum likelihood estimates for the longitudinal model parameters can be obtained by ignoring the missingness process.

4 A dynamic representation

In some cases, the hypothesis of time-constant random coefficients can be too restrictive for the longitudinal outcome and may not help model individual-specific latent dynamics (Bartolucci and Farcomeni, 2015). In this section, we introduce a dynamic specification of the RCBDM described in the previous section. As we have limited information on \( R_{it} \), we will consider time-varying random coefficients for the longitudinal data model only. The proposal can be easily generalized to deal with generic dynamic
random coefficients when needed, e.g. when modeling the time to drop-out ($T_i$) via a discrete time (parametric) survival model.

Let $Z_{it}$ denote an individual-specific, time-varying, latent variable defined on the support set $\{1, \ldots, G\}$ and let $Z_i = (Z_{i1}, \ldots, Z_{iT_i})$ be the $T_i$-dimensional latent vector associated to the $i$-th subject, $i = 1, \ldots, n$. In the following, $z_{it}$ and $z_i$ will be used to denote the generic realizations of $Z_{it}$ and $Z_i$, respectively. As before, $U_i$ denotes an individual-specific, time-constant, latent variable defined on the support set $\{1, \ldots, K\}$, while $u_i$ is the corresponding realization, with $i = 1, \ldots, n$. We assume that latent variables $Z_i$ and $U_i$ influence the longitudinal and the missing data process, respectively. In particular, conditional on $Z_i = z_i$, the generic element $Y_{it}$ of the longitudinal vector $Y_i$ depends on $Z_{it}$ only and the joint (conditional) density for the observed longitudinal sequence from the $i$-th subject is

$$f_y(y_i^o \mid Z_i = z_i) = \prod_{t=1}^{T_i} f_y(y_{it} \mid Z_{it} = z_{it}).$$

We further assume that, conditional on $U_i = u_i$, the binary random variables $R_{i1}, \ldots, R_{iT_i}$ are independent with joint (conditional) density

$$f_r(r_i \mid U_i = u_i) = \prod_{t=1}^{T_i} f_r(r_{it} \mid U_i = u_i).$$

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

$$\begin{align*}
\left\{ 
\begin{array}{l}
g[E(Y_{it} \mid Z_{it} = g)] = \zeta_g + x_{it}'\beta, \\
\text{logit}[\Pr(R_{it} = 0 \mid U_i = k)] = \xi_k + w_{it}'\gamma.
\end{array}
\right.
\end{align*}$$

In the expressions above, $g(\cdot)$ represents an appropriate link function, while the parameters $\beta$ and $\gamma$ denote the fixed effects associated of covariates $x_{it}$ and $w_{it}$, respectively. Last, $\zeta_g$ and $\xi_k$ denote the discrete random intercepts in the longitudinal and in the missing data model associated to $Z_{it} = g$ and $U_i = k$, respectively, for $g = 1, \ldots, G$.
and \( k = 1, \ldots, K \).

To model the potential dependence between \( Z_i \) and \( U_i \) and, therefore, between the longitudinal and the missing data process, let us consider a discrete upper-level latent variable \( V_i \) defined on the support set \{1, \ldots, H\}, with \( \tau_h = \Pr(V_i = h) \) for \( h = 1, \ldots, H \). We assume that, conditional on \( V_i \), the latent variables \( Z_i \) and \( U_i \) are independent, with joint distribution described by the following mixture model:

\[
f_{z,u}(Z_i, U_i) = \sum_{h=1}^{H} \tau_h [\Pr(Z_i = z_i \mid V_i = h) \Pr(U_i = u_i \mid V_i = h)].
\]

In particular, we assumed that, conditional on the \( h \)-th component of the upper-level mixture, that is \( V_i = h \), the latent variables \( Z_{it} \) evolve over time according to a first order hidden Markov chain, with initial probability vector \( \delta_h \) and transition probability matrix \( Q_h \). The corresponding elements are given by

\[
\delta_{gh|h} = \Pr(Z_{i1} = g \mid V_i = h),
\]

\[
q_{gg'|h} = \Pr(Z_{it} = g' \mid Z_{it-1} = g, V_i = h),
\]

with \( g, g' = 1, \ldots, G \) and \( h = 1, \ldots, H \). As it can be noticed, the adopted parameterization is quite complex and, this could lead to an over-parameterized model. Therefore, in order to avoid numerical difficulties in deriving maximum likelihood estimates and to reduce the number of parameters, we define \( \delta_h \) and \( Q_h \) according to the global logit parameterization suggested by Colombi and Forcina (2001).

To start, let us introduce the following set of inequality constraints on the random intercepts in the longitudinal data process:

\[
\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_G,
\]

so that lower values of \( \zeta \) (and therefore of \( Z_{it} \)) correspond to lower expected values for the longitudinal responses. Furthermore, we define the initial probabilities \( \delta_{g|h} \) as
follows:

\[
\log \frac{\Pr(Z_{it} \geq g \mid V_i = h)}{\Pr(Z_{it} < g \mid V_i = h)} = \log \frac{\delta_{g|h} + \cdots + \delta_{G|h}}{\delta_{1|h} + \cdots + \delta_{g-1|h}} = \alpha_{0g} + \psi_{0h}, \tag{5}
\]

with \( h = 1, \ldots, H \) and \( g = 2, \ldots, G \). For identifiability purposes, we set \( \psi_{01} = 0 \), so that the number of parameters to be estimated reduces from \( H(G - 1) \) to \( (G - 1) + (H - 1) \).

Similarly, the transition probabilities \( q_{gg'|h} \) are modelled as follows:

\[
\log \frac{\Pr(Z_{it} \geq g \mid \Pr(Z_{it-1} = g', V_i = h))}{\Pr(Z_{it} < g \mid \Pr(Z_{it-1} = g', V_i = h))} = \log \frac{q_{gg'|h} + \cdots + q_{Gg'|h}}{q_{1g'|h} + \cdots + q_{g-1g'|h}} = \alpha_{1gg'} + \psi_{1h}, \tag{6}
\]

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

### 4.1 Model interpretation

The modeling approach we propose offers a great flexibility. To start, unverifiable parametric assumptions on the random coefficient distribution may be avoided. Also, different individual-specific behaviors may be accommodated via this bi-dimensional latent structure. In particular, the upper-level latent variable influences both the way subjects move across the states of the hidden Markov chain and their propensity to stay into the study. Transitions between states may be more relevant for subjects belonging to a specific upper-level class and less relevant for others. Also, different propensities to dropout from the study described by the latent variable \( U_i \) may be observed for different upper-level classes. For instance, as we will see for the analysis of the Leiden 85+ data, the upper-level latent variable \( V_i \) may help identify different health conditions of subjects under observation. In the set of subjects with a better (worse) condition, we may distinguish those with a higher propensity to drop-out from the study, those with a lower propensity to drop-out, and those with complete data records. These differences may be described by the latent variable \( U_i \) in the missing data model. In a
similar fashion, the health conditions of a subject may lead to a different evolution in the longitudinal response over time. By letting the hidden Markov chain \( Z_i \) depend on the upper-level latent variable \( V_i \), non-homogeneous dynamics in the longitudinal profiles can be easily accommodated.

A further advantage of the proposed model specification is related to the chance of accounting for a potentially non-ignorable missing data process when \( H > 1 \). The independence model, leading to an ignorable missingness, is directly nested within the proposed parameterization and corresponds to \( H = 1 \). In this case, the joint density \( f_{z,u}(Z_i, U_i) \) factorizes into the product of the corresponding marginals, i.e. \( f_{z,u}(Z_i, U_i) = f_z(Z_i) f_u(U_i) \), and the individual sequences \( Y_i \) and \( R_i \) are independent. Therefore, in a sensitivity analysis perspective, \( H \) can also be interpreted as a non-ignorability parameter for the missing data process.

## 5 Parameter estimation

Estimation of model parameters can be carried out using a maximum likelihood approach. Due to the local independence assumption, the observed data likelihood is defined by

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

where \( \theta \) denotes the set of all free model parameters and \( \pi_{u_i|h} \) is used in place of \( \Pr(U_i = u_i \mid V_i = h) \) to simplify the notation. To avoid multiple summations over all possible realizations of the hidden chain, \( Z_i \), we may rely on the EM algorithm (Dempster et al., 1977). To this purpose, let \( a_{itg} = 1 \) if the \( i \)-th subject is in the \( g \)-th state at occasion \( t \), and let \( a_{itg'} = a_{it-1g} \times a_{itg} \), with \( g = 1, \ldots, G \). Similarly, let \( d_{ik} \) and \( e_{ih} \) be the indicator variables for \( U_i = k \) and \( V_i = h \), respectively. To derive
parameter estimates, we define the following complete data log-likelihood function:

\[
\ell_c(\theta) = \sum_{i=1}^{n} \left\{ \sum_{h=1}^{H} e_{ih} \log \tau_h + \sum_{h=1}^{H} \sum_{g=1}^{G} e_{ih} a_{i1g} \log \delta_{g|h} + \right.
\]

\[
+ \sum_{h=1}^{H} \sum_{t=2}^{T_i} \sum_{g=1}^{G} \sum_{g'=1}^{G} e_{ih} a_{itgg'} \log q_{gg'|h} + \sum_{t=1}^{T_i} \sum_{g=1}^{G} a_{itg} \log f_y(y_{it} | Z_{it} = g)
\]

\[
+ \sum_{h=1}^{H} \sum_{k=1}^{K} e_{ih} d_{ik} \log \pi_{k|h} \right\},
\]

(7)

The E-step of the algorithm consists in calculating the expected value of the complete data log-likelihood over the unobserved component indicators, conditional on the observed data \((y^o_i, r_i)\) and the current value of the parameter estimates \(\hat{\theta}^{(r)}\). That is,

\[
Q(\theta | \hat{\theta}^{(r)}) = \sum_{i=1}^{n} \left\{ \sum_{h=1}^{H} \hat{e}_{ih} \log \hat{\tau}_h + \sum_{h=1}^{H} \sum_{g=1}^{G} \hat{e}_{ih} \hat{a}_{i1g|h} \log \hat{\delta}_{g|h} + \right.
\]

\[
+ \sum_{h=1}^{H} \sum_{t=2}^{T_i} \sum_{g=1}^{G} \sum_{g'=1}^{G} \hat{e}_{ih} \hat{a}_{itgg'|h} \log \hat{q}_{gg'|h} + \sum_{t=1}^{T_i} \sum_{g=1}^{G} \hat{a}_{itg} \log f_y(y_{it} | Z_{it} = g)
\]

\[
+ \sum_{h=1}^{H} \sum_{k=1}^{K} \hat{e}_{ih} \hat{d}_{ik|h} \log \hat{\pi}_{k|h} \right\}
\]

(8)

where indicator variables in equation (7) are replaced by the corresponding conditional expectations. Such a computation can be consistently simplified by slightly modifying the standard forward/backward variable approach which is typically used in the hidden Markov model framework (Baum et al., 1970; Welch, 2003). To this purpose, let us define the forward and the backward variables as follows:

\[
A^{(t)}_{i|g|h} = f(y_{i1}, \ldots, y_{it}, Z_{it} = g | V_i = h),
\]

\[
B^{(t)}_{i|g|h} = f(y_{it+1}, \ldots, y_{T_i} | Z_{it} = g, V_i = h).
\]

The above quantities can be recursively derived following similar arguments to those detailed in Baum et al. (1970). Once they are computed, the posterior expectation of
the indicator variables in equation (7) is given by

$$
\hat{a}_{itg|h} = \frac{\sum_k A_{ig|h}^{(t)} B_{ig|h}^{(t)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h}{\sum_g \sum_k A_{ig|h}^{(t)} B_{ig|h}^{(t)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h},
$$

$$
\hat{\pi}_{itg'|h} = \frac{\sum_k A_{ig|h}^{(t-1)} q_{gg'|h} f(y_{it} \mid Z_{it} = g') B_{ig|h}^{(t)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h}{\sum_k \sum_g \sum_g' A_{ig|h}^{(t-1)} q_{gg'|h} f(y_{it} \mid Z_{it} = g') B_{ig|h}^{(t)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h},
$$

$$
\hat{d}_{ik|h} = \frac{\sum_g A_{ig|h}^{(T_i)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h}{\sum_k \sum_g A_{ig|h}^{(T_i-1)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h},
$$

$$
\hat{e}_{ih} = \frac{\sum_k \sum_g A_{ig|h}^{(T_i)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h}{\sum_h \sum_g \sum_k A_{ig|h}^{(T_i)} f_r(r_i \mid U_i = k) \pi_{k|h} \tau_h}.
$$

The remaining posterior probabilities can be computed as

$$
\hat{\pi}_{k|h} = \sum_h \hat{\pi}_{itg|h} \tau_h \quad \text{and} \quad \hat{\tau}_h = \frac{1}{n} \sum_{i=1}^n \hat{e}_{ih},
$$

with $k = 1, \ldots, K$ and $h = 1, \ldots, H$, respectively. To estimate the parameters for the hidden Markov chain, the following M-step equations need to be solved:

$$
\sum_{i=1}^n \sum_{h=1}^H \sum_{g=1}^G \hat{e}_{ih} \hat{a}_{i1g|h} \frac{\partial \log \delta_{g|h}}{\partial \eta_0} = 0,
$$

$$
\sum_{i=1}^n \sum_{h=1}^H \sum_{g=1}^G \sum_{g'=1}^G \hat{e}_{ih} \hat{a}_{itg'|h} \frac{\partial \log q_{gg'|h}}{\partial \eta_1} = 0,
$$

where $\eta_0 = \{\alpha_{og}, \psi_{0h}, g = 2, \ldots, G, h = 2, \ldots, H\}$ and $\eta_1 = \{\alpha_{1gg'}, \psi_{1h}, g = 1, \ldots, G, g' = 2, \ldots, G, h = 2, \ldots, H\}$.

The parameters for the longitudinal data process, $\Psi = (\beta, \zeta_1, \ldots, \zeta_G)$, are updated
by solving
\[ \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{g=1}^{G} \hat{a}_{itg} \frac{\partial \log f(y_{it} | Z_{it} = g)}{\partial \Psi} = 0 \]
under the constraints: \( \zeta_1 \leq \cdots \leq \zeta_G \). Last, for the missing data process, we need to solve:
\[ \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{k=1}^{K} \hat{d}_{ik} \frac{\partial \log f(r_{it} | U_i = k)}{\partial \Phi} = 0, \]
where \( \Phi = (\gamma, \xi_1, \ldots, \xi_K) \). The E- and the M-steps are iterated until convergence, specified in terms of the log-likelihood or the parameter values, using appropriate relative or absolute norms, e.g. \( \| \theta^{(r)} - \ell(\theta^{(r-1)}) \| < \varepsilon \) or \( \| \theta^{(r)} - \theta^{(r-1)} \| < \varepsilon \). To avoid local maxima, for a given choice of \((G, K, H)\), the algorithm is initialized from multiple starting values; at the end, the model with the highest log-likelihood value, is kept as the optimal solution.

5.1 Standard errors and model selection

Standard errors for parameter estimates obtained at convergence of the EM algorithm, \( \hat{\theta} \), can be computed using the standard sandwich formula (see e.g. White, 1980; Royall, 1986). For this purpose, we start by re-parameterizing some of the elements in \( \theta \) to obtain a vector of unconstrained parameters via the following logit transforms:
\[ \tau_h^* = \log \frac{\tau_h}{\tau_1}, \quad h = 2, \ldots, H, \]
\[ \pi_{k|h}^* = \log \frac{\pi_{k|h}}{\pi_{1|h}}, \quad k = 2, \ldots K, \quad h = 1, \ldots, H. \]

Denoting by \( \theta^* \) the vector of transformed parameter estimates, the sandwich estimate of the covariance matrix for \( \theta^* \) is defined by
\[ \text{Cov}(\hat{\theta}^*) = J(\hat{\theta}^*)^{-1} \hat{K}(\hat{\theta}^*) J(\hat{\theta}^*)^{-1}, \] (9)
where $J(\hat{\theta}^*)$ represents the observed information matrix and $\hat{K}(\hat{\theta}^*)$ provides an estimate of the covariance matrix for the score vector, defined by

$$K(\theta^*) = \text{Cov} \left[ \frac{\partial \ell(\theta^*)}{\partial \theta^*} \right] = \sum_{i=1}^{n} \text{Cov} \left[ \frac{\partial \ell_i(\theta^*)}{\partial \theta^*} \right].$$

In particular, $J(\hat{\theta}^*)$ is obtained by computing the first numerical derivative of the score vector $S(\theta^*) = \partial \ell(\theta^*)/\partial \theta^*$, evaluated at $\hat{\theta}^*$. On the other hand, $K(\theta^*)$ is estimated by $\hat{K}(\hat{\theta}^*) = \sum_{i=1}^{n} S_i(\hat{\theta}^*) S(\hat{\theta}^*)'$, where $S_i(\hat{\theta}^*)$ denotes the individual contribution to the score function for the $i$-th subject, evaluated at $\hat{\theta}^*$.

Standard errors for $\hat{\theta}^*$ are obtained as the square root of the diagonal elements in $\hat{\text{Cov}}(\theta^*)$. These can be expressed on the original scale by adopting the delta method:

$$\hat{\text{Cov}}(\hat{\theta}) = M(\theta^*) \hat{\text{Cov}}(\theta^*) M(\theta^*)',$$

where $M(\theta^*) = \partial \theta / \partial (\theta^*)'$ and, then, taking the values on the corresponding diagonal.

As it is frequent in the mixture model framework, the EM algorithm is run by treating the number of classes and states as fixed and known. The algorithm is run for varying choices for $(G, K, H)$ and the best model is chosen via model selection techniques (e.g. AIC - Akaike, 1973 or BIC - Schwarz, 1978).

### 6 Back to the Leiden 85+ study

In this section, we apply the dynamic RCBDM to the longitudinal data from the Leiden 85+ study we described in Section 2. As we highlighted before, we aim at understanding the effect of demographic and genetic conditions on the evolution of cognitive functioning in the elderly. We expect the non-homogeneous hidden Markov chain to offer a clear description of the observed longitudinal sequences and to effectively capture the dependence between longitudinal responses coming from the same subject. When compared to time-constant random coefficient models, HMMs turn out to be more flexible and provide a more concise description of the data. Similarly,
the non-homogeneous mixture for the drop-out model is expected to identify different individual propensities to stay under observation and capture dependence between missingness indicators. Last, we expect the upper-level mixture to represent a flexible and effective way to describe the potential dependence between the longitudinal responses and the missing data indicators recorded from the same subject. Since the basic independence model \((H = 1)\) is nested into the proposed specification, we may easily analyze the robustness of the estimates with respect to assumptions upon the dependence between the longitudinal response and the missingness indicators. With these objectives in mind, we start from the definition of the following models for the longitudinal response and the missing data indicator:

\[
\begin{align*}
E(Y_{it} \mid Z_{it} = g) &= \zeta_g + x_{it}'\beta \\
\text{logit}[\Pr(R_{it} = 0 \mid U_i = k)] &= \xi_k + w_{it}'\gamma,
\end{align*}
\]

where \(Y_{it} = \log[1+(30-\text{MMSE}_{it})]\) denotes the longitudinal response variable, while \(x_{it}\) and \(w_{it}\) represent the set of covariates in the two equations, respectively. In particular, the covariate set is common, that is \(x_{it} = w_{it}\), and include age (measured in terms of deviation from the age at the study entry – 85), sex (reference = female), educational status (reference = primary), and APOE genotype (reference = APOE33). It is worth to highlight that, out of the 561 subjects participating into the study, only 541 provided complete covariate information, so that the analysis is conducted on these subjects only \((i = 1, \ldots, 541)\).

To identify the optimal number of latent classes and states, we run the EM algorithm for \(G, K = 2, \ldots, 5\) and \(H = 1, 2, 3\), considering a multi-start strategy based on 50 starting values. As it is frequently done in the HMM framework, the optimal model was selected according to the BIC index using the number of observed individuals \(n\) to penalize the log-likelihood function. It is known that this is quite a conservative choice but, since model interpretability is of primary interest, this represents a reasonable choice. Values of the adopted penalized likelihood criterion are reported in Table 1.
Based on these results, we may conclude that the optimal solution corresponds $H = 2$, $G = 5$ and $K = 3$ (BIC = 5256.45). Selecting the model with $G = 5$ hidden states, we consider a solution which lies on the boundary of Table 1; however, we decided to retain it, without proceeding further with higher $G$ values, to preserve the interpretability of model parameters.

### 6.1 Results: the upper-level mixture

As stated before, the upper-level stratum of the proposed model specification allows us to describe unobserved, individual-specific, features determining the potential relation between the longitudinal and the missing data process. The selection of a model with $H = 2$ upper-level components defines a Missing Not At Random (MNAR) mechanism and suggests a potential dependence between the random effects in two equations. In particular, based on the estimated parameters, we may conclude that 21.9% of the observed subjects belongs to the first upper-level class ($\tau_1 = 0.219$), while the remaining 78.1% belongs to the second one ($\tau_2 = 0.781$). Obviously, they could be more easily interpreted if we look at the (conditional) behavior of individuals with respect to the longitudinal and the missingness process.

### 6.2 Results: the longitudinal data process

Parameter estimates for the longitudinal data model, with the corresponding standard errors, are reported in Table 2. Focusing on the estimated random intercepts, we may conclude that higher hidden states correspond to higher baseline response values, that is to individuals with worse cognitive functioning. As regards the covariate effects, estimates highlight that health conditions worsen with increasing age, while participants with higher education tend to be less cognitively impaired. This effect may be due to higher socio-economic states and to better life conditions especially during childhood. When looking at the effect of Apolipoprotein E genotype, a differential effect of $APOE_{24}$ and $APOE_{34−44}$ with respect to the baseline category ($APOE_{33}$) is observed.
Table 1: Leiden 85+ data: model selection. BIC values for varying $H, G$ and $K$

|   | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $H=1$ | $H=2$ | $H=3$ | $H=1$ | $H=2$ | $H=3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $G=2$ | 6175.95 | 6159.68 | 6162.46 | 6157.82 | 6183.51 | 6170.66 | 6175.04 | 6175.59 | 6208.19 | 6201.41 |
| $G=3$ | 5701.23 | 5684.96 | 5687.74 | 5683.10 | 5693.74 | 5679.53 | 5683.30 | 5671.47 | 5706.25 | 5713.48 |
| $G=4$ | 5491.87 | 5475.60 | 5478.38 | 5473.74 | 5480.68 | 5466.31 | 5463.99 | 5495.70 | 5487.58 | 5495.00 |
| $G=5$ | 5289.48 | 5278.05 | 5280.83 | 5293.42 | 5269.38 | 5261.76 | 5262.26 | 5276.92 | 5270.26 | 5278.42 |

Table contents: BIC values for varying $H, G$ and $K$ in Leiden 85+ data. The table shows the BIC values for different combinations of $H$, $G$, and $K$, with $H$ varying from 1 to 3, $G$ varying from 2 to 5, and $K$ varying from 2 to 5. The entries highlight the BIC values for each combination, with the lowest BIC values indicating the best model selection.
Table 2: Leiden 85+ data: parameter estimates and standard errors for the longitudinal data model

|                | Estimates | Se  |
|----------------|-----------|-----|
| $\zeta_1$      | 0.260     | 0.067|
| $\zeta_2$      | 1.134     | 0.057|
| $\zeta_3$      | 1.783     | 0.049|
| $\zeta_4$      | 2.441     | 0.054|
| $\zeta_5$      | 3.036     | 0.070|
| Age            | 0.056     | 0.007|
| High Edu       | -0.265    | 0.055|
| Male           | -0.100    | 0.055|
| APOE22−23      | 0.057     | 0.051|
| APOE24         | -0.224    | 0.049|
| APOE34−44      | 0.249     | 0.064|

However, it is worth to highlight that the estimated parameters for $APOE_{24}$ should be carefully considered as the sample is strongly unbalanced with respect to alleles, and only 12 subjects (out of 541) present such a condition.

To better understand how cognitive functioning evolves over time, we may look at the estimated parameters for the hidden Markov chain. These suggest how subjects in the $h$-th upper-level class move across hidden states over time and, therefore, we may get a clearer description of the observed longitudinal patterns. With the aim of reducing the number of estimated parameters, we considered the parametric specifications in equations (5) and (6) for $\delta_h$ and $Q_h$, with $h = 1, 2$. Such specifications lead to the results reported in Table 3. As it can be noticed, the upper-level structure helps distinguish two different behaviors in the individual sequences. The first upper-level component identifies subjects who move quite rapidly towards higher hidden states; that is, it identifies participants whose cognitive functioning gets rapidly worse when compared to the study entrance. On the other hand, the second upper-level component denotes a moderate, albeit progressive, increase in the response variable over time, with individuals presenting a slower approach to more impaired statuses.
Table 3: Leiden 85+ data - longitudinal data model: initial ($\delta_h$) and transition ($Q_h$) probabilities by upper-level components

| State | 1  | 2  | 3  | 4  | 5  |
|-------|----|----|----|----|----|
| $\delta_1$ | 0.02 | 0.08 | 0.26 | 0.37 | 0.26 |
| 1  | 0.07 | 0.76 | 0.17 | 0.00 | 0.00 |
| 2  | 0.02 | 0.37 | 0.59 | 0.00 | 0.02 |
| $Q_1$ | 3  | 0.00 | 0.01 | 0.55 | 0.43 | 0.02 |
| 4  | 0.00 | 0.00 | 0.00 | 0.88 | 0.12 |
| 5  | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| $\delta_2$ | 0.12 | 0.37 | 0.34 | 0.13 | 0.04 |
| 1  | 0.43 | 0.55 | 0.02 | 0.00 | 0.00 |
| 2  | 0.17 | 0.70 | 0.13 | 0.00 | 0.00 |
| $Q_2$ | 3  | 0.01 | 0.08 | 0.84 | 0.07 | 0.00 |
| 4  | 0.00 | 0.00 | 0.00 | 0.99 | 0.01 |
| 5  | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

6.3 Results: the missing data process

Maximum likelihood parameter estimates and the corresponding standard errors for the parameters in the missing data model are reported in Table 4.

Table 4: Leiden 85+ data: parameter estimates and standard errors in the missing data model

| Estimates | Se |
|-----------|----|
| $\xi_1$  | -15.188 | 1.326 |
| $\xi_2$  | -8.809  | 0.782 |
| $\xi_3$  | -3.394  | 0.370 |
| Age      | 2.474   | 0.268 |
| High Edu | -1.628  | 0.519 |
| Male     | 0.943   | 0.535 |
| APOE 22–23 | 0.572 | 0.669 |
| APOE 24  | -0.351  | 0.637 |
| APOE 34–44 | 1.175 | 0.523 |

By looking at the estimated random intercepts, we may first observe that individuals
Table 5: Leiden 85+ data - missing data model: class probabilities $\pi_h$ by upper-level components

| Class | $\pi_1$ | $\pi_2$ |
|-------|---------|---------|
| 1     | 0.01    | 0.68    |
| 2     | 0.42    | 0.17    |
| 3     | 0.57    | 0.15    |

in the first class are less likely to drop-out from the study. The propensity to drop-out increases when we consider individuals belonging to the second class and, more heavily, to the third one. Generally, these three latent classes may be labeled as completers, late, and early dropouts. As regards the fixed parameters, we may notice that both Age and APOE$_{34-44}$ are positively associated with the probability of early exit, while such a probability reduces with higher education. As in the longitudinal data model, gender does not seem to substantially influence the missingness indicator.

Table 5 shows the estimated probabilities for the latent $U_i$, conditional on the $h$-th upper-level latent class, that is $\pi_{k|h}$. Combining these results with those reported in Table 4, we may conclude that subjects in the first upper-level component present a higher chance to drop-out prematurely from the study after a premature and rapid decline of their cognitive functioning. This comes from an over-representation of the second and the third class in the first upper-level component of the the missing data model ($\pi_{2|1} + \pi_{3|1} = 0.99$) which also corresponds to a non-persistent transition matrix in the longitudinal data model.

A step-by-step reduction of cognitive skills is instead observed for subjects who generally stay longer under observation and that belong to the second upper-level components.

### 6.4 Sensitivity analysis

Results discussed for the MMSE data are based on the assumption that the mechanism generating the observed data is MNAR. With the aim of verifying the sensitivity of
parameter estimates to such an assumption, we present in this section the results obtained by the corresponding Missing At Random (MAR) model. By looking at the BIC values in Table 1 obtained with $H = 1$, we may observe that the optimal MAR model corresponds to $G = 5$ and $K = 3$; that is, no differences are present with respect to the MNAR counterpart when the number of states and classes is entailed. Table 6 reports parameter estimates and standard errors for both the longitudinal and the missing data model.

Table 6: Leiden 85+ data: parameter estimates and standard errors in the longitudinal and the missing data model under the MAR assumption

|                | Longitudinal | Missing |
|----------------|--------------|---------|
|                | Estimates    | Se      | Estimates | Se |
| $\zeta_1$     | 0.276        | 0.079   | -         | -  |
| $\zeta_2$     | 1.150        | 0.069   | -         | -  |
| $\zeta_3$     | 1.796        | 0.058   | -         | -  |
| $\zeta_4$     | 2.455        | 0.058   | -         | -  |
| $\zeta_5$     | 3.048        | 0.076   | -         | -  |
| $\xi_1$       | -            | -       | -16.086   | 1.863 |
| $\xi_2$       | -            | -       | -9.187    | 1.017 |
| $\xi_3$       | -            | -       | -3.307    | 0.313 |
| Age           | 0.052        | 0.008   | 2.764     | 0.401 |
| High Edu      | -0.280       | 0.071   | -2.172    | 0.756 |
| Male          | -0.083       | 0.070   | 0.554     | 0.438 |
| APOE$_{22-23}$| 0.044        | 0.055   | 0.474     | 0.408 |
| APOE$_{24}$   | -0.227       | 0.056   | -0.025    | 0.922 |
| APOE$_{34-44}$| 0.245        | 0.087   | 0.800     | 0.438 |

When comparing the above results with those reported in Sections 6.2-6.3, we may not observe substantial differences, but for the estimate of APOE$_{34-44}$ in the missing data model that turns to be not significant under the MAR assumption. This may be possibly due to the reduced amount of information which can be exploited when the longitudinal and the missing data process are considered as independent. Apart from this difference, estimated parameters for the longitudinal data process seems to be
Table 7: Leiden 85+ data - longitudinal data model: initial ($\delta$) and transition ($Q$) probabilities under the MAR assumption

| State | 1   | 2   | 3   | 4   | 5   |
|-------|-----|-----|-----|-----|-----|
| $\delta$ | 0.10 | 0.30 | 0.33 | 0.17 | 0.09 |
| 1     | 0.42 | 0.57 | 0.02 | 0.00 | 0.00 |
| 2     | 0.17 | 0.69 | 0.14 | 0.00 | 0.00 |
| $Q$   | 3   | 0.01 | 0.07 | 0.82 | 0.09 | 0.01 |
| 4     | 0.00 | 0.00 | 0.00 | 0.98 | 0.02 |
| 5     | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

quite robust to potential misspecification of the missing data mechanism and, therefore, render the results presented so far worth to be discussed.

To complete the analysis, we also report in Table 7 the estimated initial and transition probabilities for the longitudinal data model under the MAR assumption. As it can be noticed, faster declines in the individual skills may not be recovered when ignoring the missingness process. Rather, the initial and transition probabilities which are estimated under such a modeling assumption are quite close to those associated to the second upper-level component in the MNAR counterpart (see Table 3) which, as stated before, represents the most referenced category ($\tau_2 = 0.781$).

7 Conclusions

In this paper, we propose a random coefficient based hidden Markov model for longitudinal responses subject to drop-out. We consider a non-homogeneous hidden Markov chain to capture unobserved dynamics in the longitudinal data model. Similarly, a non-homogeneous finite mixture is considered for the missing data model to capture sources of unobserved heterogeneity that influence premature exits from the study. Last, an upper-level mixture is introduced to model the dependence between the random coefficients in the two profiles of interest. The proposed model specification offers great flexibility and provides a clear and concise description of the observed data. The appli-
cation to data from the Leiden 85+ study shows the strengths of the model and leads to the identification of two well distinguished sets of study participants. The former is characterized by subjects who drop-out out prematurely from the study after experiencing a severe health worsening, as it is clear from the estimated transition probability matrix. On the other hand, the second upper-level class identifies subjects who stay longer under observation experiencing moderate, albeit progressive, health worsening during the follow-up. As regards the effect of socio-demographic and genetic factors on the evolution of cognitive functioning in the elderly, the results discussed so far are in line with previous studies on the topic. As expected, age negatively influences individual skills, a higher educational level represents quite a protective factor, while $\epsilon_4$ carriers seems to present a slower decline of their cognitive functioning.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, pages 267–281. Akademiai Kiado.

Alfò, M. and Aitkin, M. (2000). Random coefficient models for binary longitudinal responses with attrition. Statistics and Computing, 10:279–287.

Alfó, M. and Maruotti, A. (2009). A selection model for longitudinal binary responses subject to non-ignorable attrition. Statistics in Medicine, 28(19):2435–2450.

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77:131–148.

Bartolucci, F. and Farcomeni, A. (2015). A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. Biometrics, 71:80–89.
Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC.

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

Bootsma-Van Der Wiel, A., Van Exel, E., De Craen, A., Gussekloo, J., Lagaay, A., Knook, D., and Westendorp, R. (2002). A high response is not essential to prevent selection bias: results from the leiden 85-plus study. *Journal of Clinical Epidemiology*, 55:1119–1125.

Colombi, R. and Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, 88:1007–1019.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M. (2010). A sensitivity analysis for shared-parameter models for incomplete longitudinal data. *Biometrical Journal*, 52:111–125.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M. (2011). Generalized shared-parameter models and missingness at random. *Statistical Modelling*, 11:279–310.

De Gruttola, V. and Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014.

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

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). “Mini-Mental State”: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12:189–198.
Gottfredson, N. C., Bauer, D. J., and Baldwin, S. A. (2014). Modeling change in the presence of nonrandomly missing data: Evaluating a shared parameter mixture model. *Structural Equation Modeling: A Multidisciplinary Journal*, 21:196–209.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, 59:829–836.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review/Revue Internationale de Statistique*, pages 221–226.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Spagnoli, A., Marino, M. F., and Alfò, M. (2017). A bi-dimensional finite mixture model for longitudinal data subject to dropout. *arXiv preprint arXiv:1707.02182*.

Tsonaka, R., Verbeke, G., and Lesaffre, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, 65(1):81–87.

Van Vliet, P., Westendorp, R. G., Van Heemst, D., De Craen, A. J., and Oleksik, A. M. (2010). Cognitive decline precedes late-life longitudinal changes in vascular risk factors. *Journal of Neurology, Neurosurgery & Psychiatry*, 81:1028–1032.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:10–13.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48:817–838.

Wu, M. C. and Bailey, K. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, 7:337–346.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44:175–188.

Zucchini, W. and MacDonald, I. (2009). *Hidden Markov models for time series*. CRC Press.