Advances in Data Analysis and Classification (2021) 15:759–793
https://doi.org/10.1007/s11634-020-00433-4

REGULAR ARTICLE

A bivariate finite mixture growth model with selection

David Aristei[1] · Silvia Bacci[2] · Francesco Bartolucci[1] · Silvia Pandolfi[1]

Received: 21 August 2019 / Revised: 30 November 2020 / Accepted: 5 December 2020 /
Published online: 29 December 2020

Abstract

A model is proposed to analyze longitudinal data where two response variables are available, one of which is a binary indicator of selection and the other is continuous and observed only if the first is equal to 1. The model also accounts for individual covariates and may be considered as a bivariate finite mixture growth model as it is based on three submodels: (i) a probit model for the selection variable; (ii) a linear model for the continuous variable; and (iii) a multinomial logit model for the class membership. To suitably address endogeneity, the first two components rely on correlated errors as in a standard selection model. The proposed approach is applied to the analysis of the dynamics of household portfolio choices based on an unbalanced panel dataset of Italian households over the 1998–2014 period. For this dataset, we identify three latent classes of households with specific investment behaviors and we assess the effect of individual characteristics on households' portfolio choices. Our empirical findings also confirm the need to jointly model risky asset market participation and the conditional portfolio share to properly analyze investment behaviors over the life-cycle.

Keywords Endogeneity · Household portfolio choices · Latent class model · Latent trajectories · Longitudinal data · Selection model

Mathematics Subject Classification 62H30

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-020-00433-4.

✉ Silvia Bacci
silvia.bacci@unifi.it

David Aristei
david.aristei@unipg.it

Francesco Bartolucci
francesco.bartolucci@unipg.it

Silvia Pandolfi
silvia.pandolfi@unipg.it

[1] Department of Economics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy

[2] Department of Statistics, Computer Science and Applications "G. Parenti", University of Florence, Viale Morgagni 59, 50134 Firenze, Italy

1 Introduction

In many contexts, longitudinal data are available where the outcome of interest, along with individual-specific covariates, is observed only conditional on a non-random selection mechanism, thus giving rise to informative missing values. For instance, two interesting situations in economics concern the time pattern of the amount of remittances from migrants to the home country (see Bacci et al. 2019) and the portfolio choices of investors over the life-cycle (see Fagereng et al. 2017). In both cases, a mechanism of selection acts generating non-random missing values: in the first case the amount of remittances is observed only when the migrant decides to send money home; in the second case, the amount of the investment is observed only when the investor is active on the financial market. In these types of context, the interest is often in clustering sample-units in homogenous groups that share a common behavior in terms of both selection variable and outcome of main interest.

In order to analyze data of the type outlined above, we propose an approach based on a bivariate latent class growth trajectory model (Muthén and Shedden 1999; Muthén 2004; Bollen and Curran 2006; Nylund et al. 2007; Bartolucci and Murphy 2015). This approach relies on a selection model component, in the sense of Heckman (1979), with a binary response variable that describes the selection phase and a continuous response variable corresponding to the outcome of main interest. Correlated error terms are also included in the model to account for the endogeneity of the selection process. Furthermore, the approach is based on the assumption that there exist latent classes (i.e., unobservable clusters defined by a discrete latent variable) of individuals with each class having a specific time trajectory for both the continuous response variable and the selection variable. Moreover, the probability of belonging to each latent class (class weight) is assumed to be affected by individual time-constant (baseline) characteristics, whereas time-varying covariates directly affect the two response variables. The resulting model we propose is thus composed of three submodels: (i) a probit model for the selection variable; (ii) a linear model for the response variable of main interest; and (iii) a multinomial logit model for the latent class membership.

As usual with latent variable models, parameter estimation is achieved through the maximum likelihood method, using an Expectation-Maximization (EM) algorithm (Dempster et al. 1977). This algorithm is based on alternating two steps that compute and maximize the expected value of the complete data log-likelihood. In order to accelerate the estimation process, after a suitable number of EM steps, the maximization of the incomplete data log-likelihood proceeds by quasi-Newton steps that directly use the score function to update the model parameters. The score vector is also used to compute, after a numerical differentiation, the observed information matrix that, in turn, allows us to obtain standard errors for the parameter estimates. The overall estimation algorithm has been implemented by means of a series of `R` functions which are available on Github at the web page https://github.com/Silvia-Pand/BivLT.

It is important to recall that the presence of the latent variable produces a model-based clustering (Fraley and Raftery 2002), with clusters corresponding to the estimated latent classes. As known, the estimation algorithm requires that the number of latent classes is specified in advance. In absence of substantial reasons that may suggest this number, its choice may be driven by information criteria typically adopted in

the finite-mixture literature, such as the Akaike Information Criterion (AIC; Akaike 1973) and the Bayesian Information Criterion (BIC; Schwarz 1978). Furthermore, once the model is estimated, the most commonly adopted approach for clustering the sample units is based on the Maximum A Posteriori (MAP) rule (Goodman 1974, 2007). According to this approach, an individual is assigned to the latent class corresponding to the highest posterior probability, that is, the conditional probability of the latent variable given the observed data.

Finally, marginal effects are computed in order to facilitate the interpretation of the regression coefficients. In particular, they are computed as the partial derivatives of the expected value of both response variables with respect to the corresponding time-varying covariates. In practice, these marginal effects allow us to evaluate how the dependent variables (outcomes of interest) change when the independent variables (covariates) change.

As an illustrative application of the proposed bivariate latent class growth trajectory model, we analyze the dynamics of portfolio choices of Italian households over the life-cycle and investigate the factors influencing the heterogeneity of both risky asset market participation and investment intensity. The empirical analysis is carried out based on an unbalanced panel dataset of Italian households from nine waves of the Bank of Italy's *Survey of Household Income and Wealth* (SHIW) over the 1998–2014 period.

Our application relies on the proposed bivariate latent class growth trajectory model that is specified in a suitable way, according to a probit submodel for the probability of participating to the financial market and a linear submodel for the share invested. Both responses are affected by time-varying socio-economic and demographic characteristics of the household. Among these time-varying covariates, an important role is played by those measuring time (i.e., year of interview and household head's age), as they drive the shape of the time trend of the response variables in the latent classes. Moreover, a multinomial logit submodel is specified for the latent class membership, being class weights dependent on time-constant household characteristics. Thus, differently from previous studies that are mainly population-average, our methodological approach allows life-cycle patterns and time trajectories of household risky investment decisions to be cluster (latent class) specific. The proposed methodological approach significantly contributes to the existing literature by allowing to explicitly take into account the existence of (unobservable) clusters of households characterized by a specific behavior in terms of both risky asset market participation and amount invested. In such a way, we are able to properly account for heterogeneity in household portfolio choices and reconcile the apparently contradictory results obtained in previous empirical studies.

In summary, the contribution of the present paper is, first of all, that of guiding the reader through using complex modeling for answering applied questions. Moreover, we also provide some methodological advances in terms of estimation with particular regard to the accelerated EM algorithm. Finally, we provide interesting results and interpretations in the specific field of application related to household risky investment decisions, also in connection with the prevailing economic theories in this field.

The remainder of the paper is organized as follows. Section 2 illustrates the proposed statistical model and its assumptions. Section 3 investigates inferential issues related

with the proposed model. In particular, we provide details on the EM algorithm used to maximize the log-likelihood function (Sect. 3.1), on the computation of the standard errors for the parameter estimates, and on some aspects related with model selection (mainly, selection of number of latent classes) and marginal effects (Sect. 3.3). Data and results of the application are described in Sect. 4, whereas in Sect. 5 we provide some final conclusions.

2 The statistical model

In this section we describe the bivariate latent growth model: we first introduce the basic notation and then we illustrate its main assumptions.

2.1 Basic notation

For a sample of n individuals, let B_{it} denote the selection variable which is equal to 1 if the continuous variable of interest, denoted by Y_{it}, is observable and to 0 otherwise, with $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$, where n is the sample size and T_i is the number of time occasions for individual i. In oder to model the informative missing mechanism, we also introduce the continuous variables B^*_{it} underlying the selection process, so that

$$\begin{aligned} B_{it} &= 1 \quad \text{if } B^*_{it} > 0, \\ B_{it} &= 0 \quad \text{if } B^*_{it} \leq 0. \end{aligned}$$

Note also that B_{it} may be unobserved for one or more occasions. This leads to a non-monotone missing patterns of the unbalanced panel data, in which an individual may not be in the sample for certain time occasions, typically because it is not interviewed by design. For instance, in the application motivating the proposed paper, B^*_{it} is the propensity of household i to participate to the risky financial market at occasion t, while Y_{it} is the percentage of investments in risky financial assets out of total financial wealth, which is held at occasion t by household i.

Let $\mathbf{B}_i = (B_{i1}, \ldots, B_{iT_i})'$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})'$ be the random vectors of binary and continuous variables previously defined for subject i. Missing observations on B_{it}, due to the absence of the unit from the sample, and consequently on Y_{it} and on the corresponding covariates, are non-informative because we rely on the *missing at random* assumption (MAR; Rubin 1976; Little and Rubin 2002) as motivated in the following.

We also denote by U_i the discrete latent variable identifying classes of individuals with the same behavior across time. The distribution of these latent variables is based on k support points, labeled from 1 to k, which correspond to the number of latent classes and have specific probabilities, as defined below. We finally denote by $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$ the observed vectors of time-varying covariates $\mathbf{W}_{it}$ and $\mathbf{X}_{it}$, affecting B_{it} and Y_{it}, respectively, and by $\mathbf{z}_i$ the observed vector of time-constant covariates $\mathbf{Z}_i$, affecting

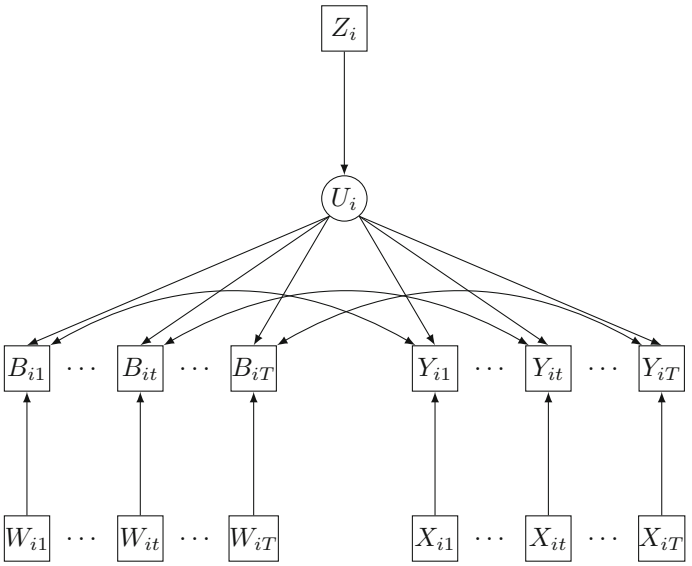

Fig. 1 Path diagram of the bivariate latent growth model, for a generic individual i

the distribution of the latent variable U_i. Vectors $\mathbf{w}_{it}$, $\mathbf{x}_{it}$, and $\mathbf{z}_i$ include a first element equal to 1 to accommodate the constant term.

Note that, since usually the main interest is in assessing the time trajectories of variables B_{it} and Y_{it}, vectors $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$ should have elements which are function of time, apart from time-varying covariates, for $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. A possible approach relies on using polynomials of order r ($r = 1, 2, \ldots$) of one or more time variables (e.g., year of interview). A common alternative consists in modeling the effect of the time through dummies for each time point (e.g., for each year of interview), but this approach is actually feasible only when the number of time points is limited. Alternatively, a semi-parametric formulation of the time effect may be based on splines (Green and Silverman 1994): this approach is more flexible with respect to the parametric one based on polynomials but it is usually less parsimonious. In the application motivating this paper, vectors $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$ include suitable polynomials both for the year of interview and the household head's age.

2.2 Model assumptions

We formulate a bivariate latent growth model (Muthén and Shedden 1999; Muthén 2004; Bollen and Curran 2006; Nylund et al. 2007) that accounts for different behaviors in the population, defined in terms of latent trajectories. A path diagram of the proposed model is displayed in Fig. 1.

Subjects are grouped into a finite number of unobservable (i.e., latent) classes characterized by homogenous behaviors. These latent classes are defined on the basis of the discrete latent variable U_i, whose distribution is given by

$$\pi_u(\mathbf{z}_i) = p(U_i = u|\mathbf{Z}_i = \mathbf{z}_i), \quad u = 1, \ldots, k.$$

The above mass probabilities, in general, depend on individual time-constant characteristics, $\mathbf{Z}_i$.

Coherently with the well-known selection model of Heckman (1979), we assume that the two responses B_{it}^* and Y_{it} have a bivariate Normal distribution, conditionally on the latent class and covariates:

$$\begin{pmatrix} B_{it}^*|U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it} \\ Y_{it}|U_i = u, \mathbf{X}_{it} = \mathbf{x}_{it} \end{pmatrix} \sim N_2[\boldsymbol{\mu}_u(\mathbf{w}_{it}, \mathbf{x}_{it}), \boldsymbol{\Sigma}], \tag{1}$$

where

$$\boldsymbol{\mu}_u(\mathbf{w}_{it}, \mathbf{x}_{it}) = \begin{pmatrix} \mathbf{w}_{it}'\boldsymbol{\beta}_u \\ \mathbf{x}_{it}'\boldsymbol{\gamma}_u \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}.$$

In the previous expressions, $\boldsymbol{\beta}_u$ is a vector of class-specific regression coefficients measuring the effect of covariates in $\mathbf{w}_{it}$, collected in matrix $\boldsymbol{\beta} = \{\boldsymbol{\beta}_u, u = 1, \ldots, k\}$, $\boldsymbol{\gamma}_u$ is a vector of class-specific regression coefficients measuring the effect of covariates in $\mathbf{x}_{it}$, collected in matrix $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}_u, u = 1, \ldots, k\}$, and ρ $(-1 \leq \rho \leq 1)$ is the correlation coefficient that accounts for the potential endogeneity of the selection. In the model, $\mathbf{x}_{it}$ is assumed to be strictly a subset of $\mathbf{w}_{it}$. Indeed, when $\mathbf{x}_{it} = \mathbf{w}_{it}$ then severe collinearity among the regressors in the two equations arises and parameters identifiability relies only on the (non-linear) functional form of the distribution (Puhani 2000). In order to alleviate these problems, as in empirical applications, exclusion restrictions are imposed according to which extra regressions are included in the selection equation for B_{it} and do not appear in the outcome equation for Y_{it} (Marchenko and Genton 2012).

It is worth noting that the proposed model differs from the selection model of Heckman (1979) for the presence of mixture components that properly account for heterogeneity in the population. It also differs from the mixture latent growth model of Bartolucci and Murphy (2015) for the introduction of the correlation term. Moreover, latter has some other specific differences driven by the particular type of application in sport dealt with. Indeed, the special case with $k = 1$ coincides with the model of Heckman (1979) and the special case with $\rho = 0$ coincides with the model of Bartolucci and Murphy (2015).

Assumption (1) implies two main correlated equations. The first equation accounts for the unobservable nature of B_{it}^* through a probit model for the probability of observing a response, that is,

$$p(B_{it} = 1|U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it}) = \Phi(\mathbf{w}_{it}'\boldsymbol{\beta}_u), \tag{2}$$

with $\Phi(\cdot)$ being the cumulative probability function of a normal distribution. The second equation is based on the assumption of normality of the response variable Y_{it} with constant variance σ^2 and expected value given by

$$E(Y_{it}|U_i = u, \mathbf{X}_{it} = \mathbf{x}_{it}) = \mathbf{x}_{it}'\boldsymbol{\gamma}_u. \tag{3}$$

We recall that Y_{it} is observed only if $B_{it} = 1$.

A multinomial logit model is also introduced to account for the effect of the individual time-constant covariates on the class membership:

$$\log \frac{\pi_u(\mathbf{z}_i)}{\pi_1(\mathbf{z}_i)} = \log \frac{p(U_i = u|\mathbf{Z}_i = \mathbf{z}_i)}{p(U_i = 1|\mathbf{Z}_i = \mathbf{z}_i)} = \mathbf{z}_i' \boldsymbol{\delta}_u, \tag{4}$$

where $\boldsymbol{\delta}_u$ is the vector of regression coefficients measuring the effect of time-constant covariates on the odds ratio of Class u against Class 1 with $u = 2, \ldots, k$. These parameters are collected in matrix $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_u, u = 2, \ldots, k\}$.

As mentioned above, in the presence of non-monotone non-informative missing observations for variable B_{it}, due to the absence from the sample of unit i at occasion t, we rely on the MAR assumption. Under this assumption, the probability of the realized missing pattern, given the observed and the unobserved data, does not depend on the unobserved data. Therefore, provided that the model for this type of missing data mechanism is separated from the proposed model, these missing responses are ignorable for likelihood based inference. The resulting model may be formulated by introducing the missing data indicator M_{it} that is equal to 1 when subject i does not answer at all at occasion t and to 0 otherwise. Thus, for a certain subject i, we collect these variables in vector $\mathbf{M}_i = (M_{i1}, \ldots, M_{iT_i})'$. The corresponding response pattern is given by $(\mathbf{m}_i, \mathbf{b}_{i,obs}, \mathbf{y}_{i,obs})$, with $\mathbf{m}_i$ being a realization of $\mathbf{M}_i$ and $\mathbf{b}_{i,obs}$ and $\mathbf{y}_{i,obs}$ being subvectors containing the observed components of $\mathbf{B}_i$ and $\mathbf{Y}_i$, respectively. We also introduce $\mathbf{W}_{i,obs}$ and $\mathbf{X}_{i,obs}$ to denote the matrices of all observed covariates for subject i.

The MAR assumption implies that the parameters of interest can be estimated on the basis of the log-likelihood of the vectors of the observed outcomes $(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs})$ only, without the model specification for non-informative missingness. In particular, based on the assumptions formulated above, the distribution of interest is as follows:

$$\begin{aligned}
&p(\mathbf{b}_{i,\,obs}, \mathbf{y}_{i,\,obs} \mid u, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}) \\
&\quad = p(\mathbf{b}_{i,\,obs}, \mathbf{y}_{i,\,obs}|U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it}, \mathbf{X}_{it} = \mathbf{x}_{it},\ t = 1, \ldots, T_i : m_{it} = 0) \\
&\quad = \prod_{t:m_{it}=0}^{T_i} p(b_{it}|u, \mathbf{w}_{it})^{1-b_{it}} f(b_{it}, y_{it}|u, \mathbf{w}_{it}, \mathbf{x}_{it})^{b_{it}},
\end{aligned}$$

where in the second expression the conditioning is on the observed covariates. Moreover, $p(b_{it}|u, \mathbf{w}_{it})$ is defined according to (2) and $f(b_{it}, y_{it}|u, \mathbf{w}_{it}, \mathbf{x}_{it})$ is the joint density of b_{it} and y_{it} based on assumption (1); the previous product is defined only for those occasions t for which the answer of subject i is observed. In particular, we need this density for $b_{it} = 1$ when it is equal to

$$\int_0^\infty \phi_2[(b_{it}^*, y_{it})', \boldsymbol{\mu}_u, \boldsymbol{\Sigma}]db_{it}^*,$$

where $\phi_2[\cdot]$ is the density function of the bivariate Normal distribution in (1).

The manifest distribution of the proposed bivariate mixture growth model is expressed as follows:

$$p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}, \mathbf{z}_i) = \sum_{u=1}^{k} \pi_u(\mathbf{z}_i)\ p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid u, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}). \quad (5)$$

This expression is crucial for inference as we explain in the following section. Another quantity of interest is the posterior probability that a subject with observed response configuration $(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs})$ belongs to latent class u. Using standard rules, the posterior probabilities are equal to

$$p(U_i = u \mid \mathbf{b}_{i,obs}, \mathbf{y}_{i,obs}, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}, \mathbf{z}_i) = \frac{\pi_u(\mathbf{z}_i)\ p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid u, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs})}{p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs})}, \quad u = 1, \ldots, k. \quad (6)$$

These probabilities are used to allocate subjects to the different latent classes, as will be clarified in the sequel.

3 Model inference

In the following we first illustrate the model estimation process, based on the maximization of the log-likelihood function. Then, we describe how to compute standard errors, selecting the number of latent classes, and assigning sample units to the latent classes. Finally, we outline how to compute marginal effects.

3.1 Maximum likelihood estimation

Given a sample of n independent units, the log-likelihood of the proposed model is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}, \mathbf{z}_i),$$

where $\boldsymbol{\theta}$ is the vector of the free model parameters, that is, $\boldsymbol{\theta} = (\boldsymbol{\beta}'_u, \boldsymbol{\gamma}'_u, \boldsymbol{\delta}'_u, \sigma^2, \rho)'$ and $p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}, \mathbf{z}_i)$ is the manifest distribution defined in (5). Note that the number k of mixture components is not included in the vector of model parameters because it has to be a priori fixed, as clarified in Sect. 3.2. In order to maximize $\ell(\boldsymbol{\theta})$, we rely on the EM algorithm Dempster et al. (1977).

The maximization algorithm is based on the *complete-data log-likelihood* that we could compute if we knew the value of the latent variable U_i for every unit i in the sample. It is defined as follows:

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{u=1}^{k} a_{iu} \log\left[\pi_u(\mathbf{z}_i)\ p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid u, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs})\right],$$

where a_{iu} is an indicator variable equal to 1 if subject i belongs to cluster u and to 0 otherwise.

As usual, the EM algorithm alternates the following two steps until convergence:

- **E-step**: it consists in computing the conditional expected value of the complete data log-likelihood given the observed data and the current value of the model parameters.
- **M-step**: it consists in maximizing the expected value of the complete data log-likelihood resulting from the E-step with respect to $\boldsymbol{\theta}$, so as to update the parameters.

In practice, at the E-step we need to compute the posterior expected value of every indicator variable a_{iu}, that is,

$$\hat{a}_{iu} = p(U_i = u \mid \mathbf{b}_{i,obs}, \mathbf{y}_{i,obs}, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}, \mathbf{z}_i), \quad u = 1, \ldots, k, \tag{7}$$

for $i = 1, ..., n$ according to (6). This value is directly used to update the parameters in $\boldsymbol{\theta}$ at the M-step, by maximizing

$$\sum_{i=1}^{n}\sum_{u=1}^{k} \hat{a}_{iu} \log p(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs} \mid u, \mathbf{W}_{i,obs}, \mathbf{X}_{i,obs}),$$

with respect to parameter vector $\boldsymbol{\beta}_u$, $\boldsymbol{\gamma}_u$, σ^2, and ρ, and by maximizing

$$\sum_{i=1}^{n}\sum_{u=1}^{k} \hat{a}_{iu} \log \pi_u(\mathbf{z}_i)$$

with respect to the parameter vectors $\boldsymbol{\delta}_u$. These optimizations are performed on the basis of suitable numerical algorithms.

The convergence of the EM algorithm is checked on the basis of the relative log-likelihood difference, that is,

$$\left[\ell(\boldsymbol{\theta}^{(s)}) - \ell(\boldsymbol{\theta}^{(s-1)})\right] / \mid \ell(\boldsymbol{\theta}^{(s-1)}) \mid < \epsilon \tag{8}$$

where $\boldsymbol{\theta}^{(s)}$ is the parameter estimate obtained at the end of the s-th M-step and ϵ is a suitable tolerance level (e.g., 10^{-8}).

In order to speed up the estimation process, after a suitable number of EM steps we run a Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method (see Givens and Hoeting 2013, and reference therein) to directly maximize the incomplete data log-likelihood, which relies on the score vector to update model parameters. The score vector is computed as the first derivative of the conditional expected value of

the complete data log-likelihood given the observe data, which has been proved to correspond to the score vector for the observed data (or incomplete data) log-likelihood (Oakes 1999). The number of EM steps performed before starting to run these steps is again driven by the relative log-likelihood difference in (8) based on a different tolerance level ϵ^* that must be defined in advance. To explore the performance, in terms of computational efficiency, of the proposed approach with respect to the classical EM algorithm we set up a small simulation study, where we run, under different scenarios, the algorithm based on different tolerance levels ϵ^* for moving to the acceleration steps. Details are provided in "Appendix A". We also evaluate the competing algorithms in our application, obtaining that the classical EM algorithm is between 2.5 and 6.5 times slower than the proposed accelerated version.

It is important to recall that the EM algorithm requires to be initialized by choosing suitable starting values for the parameters in $\boldsymbol{\theta}$. In fact, a typical problem in estimating discrete latent variable models is the multimodality of the log-likelihood function. In order to prevent this problem, we rely on a multi-start strategy, based on both a deterministic and a random rule, the latter repeated a given number of times, so as to properly explore the parameter space. Then, for a given k, we take as final parameter estimate the one corresponding to the largest log-likelihood value found at convergence.

The deterministic initialization of the algorithm consists in computing the starting values of the parameters affecting both the probability of observing a response and the response variable itself on the basis of descriptive statistics (mean and quantiles) of the observed outcomes. The starting values for the mass probabilities $\pi_u(\mathbf{z}_i)$ are chosen as $1/k$, for $i = 1, \ldots, n$ and $u = 1, \ldots, k$.

The random starting rule is instead based on random values generated from a standard normal distribution for the parameters $\boldsymbol{\beta}_u$ and $\boldsymbol{\gamma}_u$ and from a uniform distribution for parameters σ^2 and ρ. Moreover, we draw the initial values of the mass probabilities from a uniform distribution between 0 and 1 and then we normalize these random draws so that they sum to 1.

3.2 Standard errors, model selection, and clustering

After the model is estimated with a given number of classes, we obtain standard errors for the parameter estimates on the basis of the observed information matrix $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$. In particular, the standard error for each parameter is obtained as the square root of the corresponding diagonal element of the inverse of this matrix, $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$. In our application, the computation of the observed information matrix is based on a numerical method (Bartolucci and Farcomeni 2009), where $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is obtained as minus the numerical derivative of the score vector at convergence. As discussed above about the EM acceleration, the score vector is obtained analytically as the first derivative of the conditional expected value of the complete data log-likelihood, which is based on the expected frequencies $\hat{a}_{iu}$ corresponding to the final parameter estimate $\hat{\boldsymbol{\theta}}$ (Oakes 1999).

It is already clear that the number k of latent classes does not belong to the vector of free parameters $\boldsymbol{\theta}$. In fact, k has to be chosen before performing the estimation process:

its value may be suggested by substantial reasons or, alternatively, its choice may be driven by information criteria common to the finite-mixture literature (McLachlan and Peel 2000), which rely on penalized measures of model fit. In particular, to select the number of latent classes the AIC (Akaike 1973) and the BIC (Schwarz 1978) are based on the indices

$$\begin{aligned} \text{AIC} &= -2\hat{\ell} + 2\ \#\text{par}, \\ \text{BIC} &= -2\hat{\ell} + \log(n)\ \#\text{par}, \end{aligned}$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood of the model of interest and #par denotes the number of free parameters.

In practice, a series of models is estimated for increasing values of k until the value of the index corresponding to the preferred information criterion does not start to increase; then, the previous value of k is adopted as the optimal one. Following the main stream of the literature (for a review see Bacci et al. 2014, and references therein), if the two criteria lead to selecting a different number of classes, we suggest to rely on the BIC, which tends to have good performance in several contexts and is more parsimonious with respect to the AIC.

A debated issue in the LC literature concerns the selection of k when covariates affect the class membership probabilities. According to a commonly accepted recommendation (Nylund-Gibson and Masyn 2016), k should be selected relying on a model without covariates, thus avoiding to overextract classes due to the noise present in a more complex model; once the value of k is selected, covariates are then included. Unfortunately, this procedure cannot be directly applied to the bivariate latent growth model here proposed, because its equations (2) and (3), for B_{it} and Y_{it}, in addition to the equation for the class weights (4), are affected by covariates and, most of all, must differentiate for at least one regressor. This is requested by the exclusion restriction condition characterizing Heckman-type models, as clarified in Sect. 2.2. For this reason, in what follows we adopt a different strategy accounting for the relevance of the covariates (mainly, those related to the time) for our analysis. We first explore the time trajectories under basic alternative model specifications enclosing time-varying and time-constant covariates, that is, the standard Heckman model and the latent growth model with $k = 1$ and with polynomials of different orders for age and year (see Sect. 4.2). Once the order of the polynomials for both age and year has been chosen, we select k.

It is worth remarking that the choices of r (order of polynomials for age and year) and k (number of latent classes) are not unrelated and, therefore, they should be simultaneously selected by a one-step strategy. In principle, the optimal number of r and k chosen on the basis of the proposed hierarchical strategy (based on selecting first r and then k given r) might differ from the one obtained with the simultaneous selection. However, the latter one is considerably slower and, at least in the specific application here discussed, does not provide noteworthy differences (see Sect. 4.3).

An additional relevant issue when dealing with the proposed model concerns the assignment of the units to the latent classes. As usual, the estimation algorithm directly provides the estimated posterior probabilities of U_i, as defined in (7), which may be

used for this assignment. In particular, a subject is assigned to one of the k latent classes according to the standard MAP rule (or modal assignment), see Goodman (2007), which consists in allocating subject i to latent class u when $\hat{a}_{iu} = \hat{a}_i^*$, where $\hat{a}_i^*$ is the maximum of $\hat{a}_{i1}, \ldots, \hat{a}_{ik}$. Note that this phase is error prone; however the classification error resulting from the MAP assignment may be estimated using simple probability calculus (for details see Vermunt 2010). Moreover, several studies proved the MAP allocation to be superior in terms of classification error with respect to alternative methods, among which the method of the expected proportions (Goodman 2007), the method of bagging based on bootstrap (Dias and Vermunt 2008), and the one proposed by Bandeen-Roche et al. (1997) based on multiple pseudo-class draws that randomly assign individuals to latent classes for a repeated number of times according to the posterior probabilities (Bray et al. 2015).

3.3 Marginal effects

In order to favor the interpretation of the regression coefficients, we suggest to obtain the marginal effects of each covariate on the two response variables. In the case of the time-varying covariates collected in $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$, the marginal effects for individual i and occasion t are computed as follows:

$$\begin{aligned}
\frac{\partial E\left(B_{it} \mid U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it}, \mathbf{Z}_i = \mathbf{z}_i\right)}{\partial w_{itj}} &= \sum_{u=1}^{k} \hat{\pi}_u(\mathbf{z}_i) \frac{\partial E\left(B_{it} \mid U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it}\right)}{\partial w_{itj}} \\
&= \sum_{u=1}^{k} \hat{\pi}_u(\mathbf{z}_i) \phi(\mathbf{w}_{it}' \hat{\boldsymbol{\beta}}_u) \hat{\beta}_{uj}, \\
\frac{\partial E\left(Y_{it} \mid U_i = u, \mathbf{X}_{it} = \mathbf{x}_{it}, \mathbf{Z}_i = \mathbf{z}_i\right)}{\partial x_{itj}} &= \sum_{u=1}^{k} \hat{\pi}_u(\mathbf{z}_i) \frac{\partial E\left(Y_{it} \mid U_i = u, \mathbf{X}_{it} = \mathbf{x}_{it}\right)}{\partial x_{itj}} \\
&= \sum_{u=1}^{k} \hat{\pi}_u(\mathbf{z}_i)\ \hat{\gamma}_{uj}, \qquad (9)
\end{aligned}$$

where w_{itj} and x_{itj} denote specific elements of $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$. With reference to the time-constant covariates $\mathbf{z}_i$, the marginal effects are obtained as:

$$\begin{aligned}
\frac{\partial E\left(B_{it} \mid U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it}, \mathbf{Z}_i = \mathbf{z}_i\right)}{\partial z_{ij}} &= \sum_{u=1}^{k} \frac{\partial \pi_u(\mathbf{z}_i)}{\partial z_{ij}} E\left(B_{it} \mid U_i = u, \mathbf{W}_{it} = \mathbf{w}_{it}\right) \\
&= \sum_{u=1}^{k} \left\{ \hat{\pi}_u(\mathbf{z}_i) \left[\hat{\delta}_{uj} - \sum_{v=1}^{k} \hat{\pi}_v(\mathbf{z}_i) \hat{\delta}_{vj} \right] \Phi(\mathbf{w}_{it}' \hat{\boldsymbol{\beta}}_u) \right\}, \\
\frac{\partial E\left(Y_{it} \mid U_i = u, \mathbf{X}_{it} = \mathbf{x}_{it}, \mathbf{Z}_i = \mathbf{z}_i\right)}{\partial z_{ij}} &= \sum_{u=1}^{k} \frac{\partial \pi_u(\mathbf{z}_i)}{\partial z_{ij}} E\left(Y_{it} \mid U_i = u, \mathbf{X}_{it} = \mathbf{x}_{it}\right) \\
&= \sum_{u=1}^{k} \left\{ \hat{\pi}_u(\mathbf{z}_i) \left[\hat{\delta}_{uj} - \sum_{v=1}^{k} \hat{\pi}_v(\mathbf{z}_i) \hat{\delta}_{vj} \right] \mathbf{x}_{it}' \hat{\boldsymbol{\gamma}}_u \right\}. \qquad (10)
\end{aligned}$$

Accordingly, the averaged marginal effect may be computed as the overall mean of the individual marginal effects. Finally, we obtain standard errors for these marginal effects through a (non-parametric) bootstrap approach, resampling from original data a certain number of times (Efron and Tibshirani 1993).

4 Application

In this section we first illustrate the empirical background of the proposed application and describe the data. Then, we illustrate the specification of the bivariate latent growth model of household portfolio choices and we discuss the results of the data analysis. We pay specific attention to the interpretation of the estimated class-specific age and time trajectories of market participation and risky asset share, and also to the discussion of the effects of time-varying and time-constant covariates.

In "Appendix B", we provide an example of the `R` code to specify the bivariate latent growth model and display the model parameter estimates.

4.1 Data description

The standard reference for the economic theory related to households' participation in the risky asset market is the Merton portfolio selection model (Merton 1969). One of the main implications of this model is that all investors, independently of their wealth and attitudes toward risk, should participate in all risky asset markets and should hold the same fully diversified portfolio of risky securities (Guiso and Sodini 2013). However, empirical evidence on household portfolios seems to depart from these predictions. On one side, a substantial fraction of households do not participate in risky asset markets, mainly due to fixed entry or participation costs (Haliassos 2008), limited cognitive skills (Christelis et al. 2010), low level of financial literacy and education (van Rooij et al. 2011), poor health status (Edwards 2008; Atella et al. 2012), and risk aversion (Guiso and Paiella 2008). On the other side, evidence about the life-cycle pattern of the conditional risky asset share is quite controversial, having age profiles of the invested amounts been found both relatively or extremely flat (Guiso et al. 2002; Ameriks and Zeldes 2004), monotonically increasing (Alessie et al. 2004), and also monotonically decreasing (Fagereng et al. 2017).

The empirical analysis is based on micro-data from nine waves of the Bank of Italy's *Survey of Household Income and Wealth* (SHIW) over the period 1998–2014. This survey, which started in the 1960s and is carried out on a biennial basis since 1998, provides detailed information on income, wealth, consumption expenditures, and portfolio choices, as well as on household composition, demographic characteristics, and labor force participation, for a representative sample of about 8000 Italian households in each wave. In 1989 the Bank of Italy introduced a longitudinal component into the survey and, since then, an increasingly fraction of the respondents have been interviewed for two or more consecutive surveys; currently, about one half of the sample is included in the panel (see Brandolini 1999; Bank of Italy 2015, for more details on the panel structure of the SHIW).

Table 1 Percentage of households owning risky financial assets and conditional shares invested, by year

	1998	2000	2002	2004	2006	2008	2010	2012	2014
Risky asset market participation	30.32	35.49	31.80	30.93	28.76	24.77	32.81	32.64	29.78
Conditional investment share	45.20	43.32	41.71	43.74	41.35	37.26	48.16	46.99	43.77

For the aims of our analysis, we exploit the longitudinal dimension of the SHIW and define our data sample on those households that were interviewed for at least four consecutive waves. Coherently with previous empirical studies (Alessie et al. 2004; Ameriks and Zeldes 2004), this choice allows us to track household portfolio choices over a period of at least eight years, which is adequate to properly model investment dynamics while keeping the number of households sufficiently large. Moreover, we focus on households whose head is aged between 25 and 85 and, as in Guiso and Jappelli (2002), we drop observations with inconsistent responses for age, gender, and education. After this data cleaning procedure, we dispose of an unbalanced sample of 18,106 observations on 3157 households, 373 of which were interviewed in all the nine waves between 1998 and 2014.

Exploiting the detailed breakdown of household financial portfolios provided by the SHIW, we distinguish between risky and safe financial assets. In particular, following Guiso and Jappelli (2002), we define risky financial assets as the sum of directly held stocks, long-term government bonds, other bonds, mutual funds, managed investment accounts, foreign assets, and defined-contribution pension plans. The remaining assets (transaction accounts and certificates of deposit, treasury bills, and the cash value of life insurance) are classified as risk-free.

Table 1 reports the percentage of households owning risky financial assets and the shares invested in risky assets out of total financial wealth (conditionally on owning risky assets), for each year. The data in Table 1 suggest that the total participation is fairly constant over time and about 30% of the households invest in risky financial assets each year, decreasing to 28.8% and 24.8% in 2006 and 2008, respectively. We also notice that the risky asset share is constant over time and amounts to about 45% of household total financial wealth in each year (with the exception of 2008, when it reduces to 37%).

Figure 2 shows the life-cycle profiles of risky financial market participation (left panel) and share invested (right panel) for selected cohorts defined on the basis of the household head year of birth. Cohorts are defined on 5-year intervals, with the first cohort including households with head born between 1968 and 1972 (and was aged between 26 and 30 in 1998, the first survey year), and are followed (with the exception of the last two cohorts) over a 16-year period. The graphical analysis of the left panel of Fig. 2 suggests that cohort effects are likely to play an important role, as participation rates differ across cohorts observed at the same age, with successive cohorts having higher participation rates in the first part of the life-cycle and lower rates in later stages. Moreover, looking at the right panel of Fig. 2, we notice again

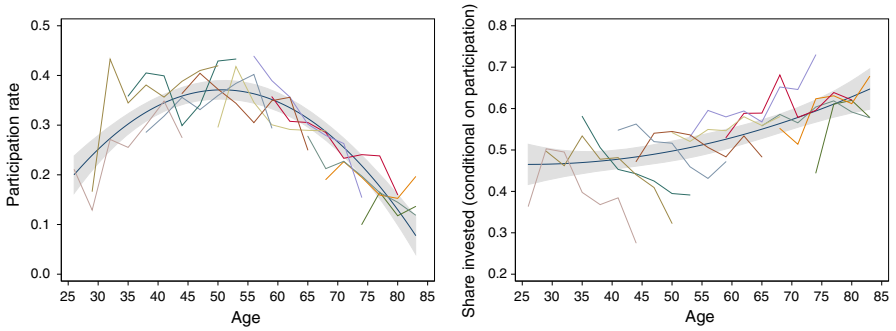

Fig. 2 Risky asset market participation (left panel) and conditional shares invested (right panel), by age

cohort-specific patterns with an overall pattern that tend to increase with age (i.e., older households invest a relatively larger share of their financial wealth in risky assets).

The evidence based on the descriptive statistics commented above suggests the existence of significant life-cycle patterns for both risky asset market participation and conditional investment shares. However, as discussed in Ameriks and Zeldes (2004), it does not allow to properly disentangle time, age, and cohort effects. In the next sections, we illustrate the results obtained with the bivariate latent class growth trajectory model illustrated in Sects. 2 and 3.

4.2 Model specification

The model is specified according to the description provided in Sect. 2.1, being B_{it}^* the propensity of household i to participate to the risky financial market at occasion t and Y_{it} the percentage of investments in risky financial assets out of total financial wealth.

In order to assess the life-cycle and time patterns of the response variables in each latent class, a polynomial for the household head's age and another polynomial for the year of interview are introduced in both vectors $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$.

Furthermore, both participation and outcome equations control for the following time-varying covariates: household disposable income (net of financial income) (`disposable income`, in thousands of euros), whether household has any debt (dummy `debts`), number of household members (`household size`), presence of children under 14 years (dummy `children`), marital status (dummy `married`), and employment status of the household head (dummies `employee` and `retired`). As common practice in estimating selection models, in order to improve model identifiability we impose an exclusion restriction and assume that asset market participation probability is also affected by the stock of real assets (`real assets`, in thousands of euros) owned by the household, by the regional unemployment rate (`unemployment rate`), and by the average number of bank branches (per 100,000 inhabitants) at regional level (`bank branch density`).

As concerns time-constant covariates affecting latent class membership, we include in vectors $\mathbf{z}_i$ the household head's gender (dummy `female`) and the values observed

at the first available time occasion for the area of residence (dummies `north` and `centre`), town size (dummy `small town`), and household head's educational level (dummies `lower secondary education`, `upper secondary education`, and `tertiary education`).

It is worth noting that, as age, year of interview, and year of birth are linearly related (i.e., *year of interview* = *age* + *year of birth*), some restrictions are necessary to properly model life-cycle patterns for both risky asset market participation and share invested (for a discussion see Ameriks and Zeldes 2004). To avoid this type of multicollinearity and to identify age, year of interview, and cohort effects, several strategies were proposed. Here, following Giuliano and Spilimbergo (2013) and Fagereng et al. (2017), we control for unrestricted time effects and proxy cohort effects by means of an exogenous variable capturing stock market returns during the household head's youth, assuming that early experiences have enduring effects on risk preferences and affect stock market participation decisions. Specifically, we use a composite indicator of stock market returns (`youth stock return` in the following), defined as a weighted average of the Italian Stock Exchange (80%) and the MSCI World Index (20%), experienced when the head was aged between 18 and 25. As this composite indicator is time-constant, we include it in vector $\mathbf{z}_i$. However, it is worth noting that its inclusion in vectors $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$ does not modify in a sensible way the results of the estimation process (results not shown here).

4.3 Model selection and latent class characterization

As preliminary and explorative analysis, we consider estimates from an Heckman (1979) model as well as a bivariate latent growth model with $k = 1$, both of them for increasing values of the order r of polynomials for age and year. For the sake of completeness, a less parametric version of the bivariate latent growth model is estimated where the polynomial for the survey year is replaced by time dummies. Table 2 shows a summary of the main results for each estimated model: maximum log-likelihood, value of BIC, estimated value of correlation coefficient between probability of investing and share invested, and the variance parameter, together with the corresponding standard errors.

As expected, the results based on the Heckman (1979) model are the same as those of our proposed model with $k = 1$. Moreover, we first observe that all models agree on the presence of a statistically significant negative correlation between the probability of investing in risky assets and the share invested. Second, the BIC values lead to the selection of order $r = 4$ for the polynomial of age and, at the same time, they outline that the better fit of models with dummies for survey year is not sufficient to offset the loss of parsimony. Anyway, we verified that the choice between polynomial and dummies for variable year does not significantly affect the parameter estimates.

In light of these results, we base our analysis on a latent growth model with correlated components, specified as in Eqs. (2)–(4), and with two polynomials of order $r = 4$ both for the household head's age and the year of interview.

As far as the choice of the number k of mixture components, the selection procedure is based on the BIC, as illustrated in Sect. 3.2. In particular, the sequence of latent

Table 2 Explorative analysis: comparison between basic models

Model	Polynomial for age	Polynomial for year	$\hat{\ell}$	BIC	$\hat{\rho}$	s.e. $\hat{\rho}$	$\hat{\sigma}^2$	s.e. $\hat{\sigma}^2$
Heckman	$r = 3$	$r = 3$	$-10,010.99$	20,287.87	-0.239	0.068	0.093	0.004
Heckman	$\mathbf{r = 4}$	$\mathbf{r = 4}$	-9984.13	**20,266.37**	-0.244	0.068	0.093	0.004
Heckman	$r = 5$	$r = 5$	-9975.44	20,281.23	-0.250	0.068	0.093	0.004
Latent growth ($k = 1$)	$r = 3$	$r = 3$	$-10,011.32$	20,288.53	-0.241	0.047	0.093	0.002
Latent growth ($k = 1$)	$\mathbf{r = 4}$	$\mathbf{r = 4}$	$--9984.47$	**20,267.06**	-0.246	0.047	0.093	0.002
Latent growth ($k = 1$)	$r = 5$	$r = 5$	-9975.80	20,281.94	-0.252	0.046	0.093	0.002
Latent growth ($k = 1$)	$r = 3$	Dummies	-9972.04	20,290.55	-0.252	0.047	0.093	0.002
Latent growth ($k = 1$)	$\mathbf{r = 4}$	**Dummies**	-9955.27	**20,273.11**	-0.251	0.046	0.093	0.002
Latent growth ($k = 1$)	$r = 5$	Dummies	-9954.73	20,288.15	-0.251	0.046	0.093	0.002

The table reports log-likelihood ($\hat{\ell}$), BIC index, estimated correlation coefficient ($\hat{\rho}$) and variance ($\hat{\sigma}^2$) for the Heckman sample selection model and for the bivariate latent growth model with $k = 1$

The minimum BIC value for each type of model is reported in bold

Table 3 Selection of the number of mixture components

k	$\hat{\ell}$	BIC	$\hat{\rho}$	s.e. ($\hat{\rho}$)	BIC ($\rho = 0$)
1	−9984.47	20,267.06	−0.246	0.047	20,295.50
2	−8648.46	17,965.68	−0.220	0.063	17,987.27
3	−8290.64	**17,620.68**	−0.115	0.090	17,636.96
4	−8129.68	17,669.40	−0.070	0.104	17,683.36

The table reports log-likelihood ($\hat{\ell}$), BIC index, estimated correlation coefficient ($\hat{\rho}$) and related standard error, BIC index for the special case of uncorrelated components ($\rho = 0$), for $k = 1, 2, 3, 4$. The minimum BIC value is reported in bold

growth models including the set of covariates mentioned in Sect. 4.2 and polynomials of order four for age and year, provides the values of the BIC index shown in Table 3. The BIC values for the special case of $\rho = 0$ are also displayed in the last column of the table.

Accordingly, we adopt a model with $k = 3$, corresponding to the minimum value of BIC. It is also worth noting that the estimated correlation coefficient $\hat{\rho}$ is negative and decreasing in absolute value while its standard error increases, for k ranging from 1 to 4. In particular, for $k = 3$ (and $k = 4$) the correlation coefficient is not statistically significant. To provide evidence of the robustness of results discussed in the following, the bivariate latent growth model with $k = 3$ mixture components and with $\rho = 0$ was also estimated, and no relevant difference resulted in the main conclusions.

Moreover, as an additional robustness check, we also selected r and k by adopting a one-step strategy, which led to choose $k = 4$ and $r = 3$. Despite the differences in the values of k and r, this model (results here omitted for the sake of space) presents several similarities with the one presented in the paper (having $k = 3$ and $r = 4$): in particular, estimates of ρ and σ^2 are very close to each other, and two latent classes have similar profiles in both models.

From the results obtained under the selected model, the class collecting the main part of households is Class 2 with an average mass probability $\hat{\bar{\pi}}_2$ equal to 0.498, followed by Class 3 with average weight $\hat{\bar{\pi}}_3$ equal to 0.333, whereas the smallest class is the first with average weight $\hat{\bar{\pi}}_1$ equal to 0.169, where we define $\hat{\bar{\pi}}_u = \sum_i \pi_u(\mathbf{z}_i)/n$, $u = 1, \ldots, k$.

To characterize the three latent classes, we allocate each household to these classes on the basis of the posterior probabilities, estimated as in (7), which account for both the observed pattern of response variables $(\mathbf{b}_{i,obs}, \mathbf{y}_{i,obs})$ and the prior probabilities $\pi_u(\mathbf{z}_i)$. As reported in Table 4, the 16.4% of households is allocated in Class 1, the 54.1% in Class 2, and the remaining 29.5% in Class 3.

Table 4 also shows the average values of time-varying and time-constant covariates for each latent class. Moving from Class 2 to Class 1 through Class 3, we observe increasing average values of household disposable income as well as of real assets: Class 1 clearly emerges as the wealthiest group, both in terms of annual income flows and of real assets possessed. From Class 2 to Class 1, we also note an increasing proportion of households living in the North, and with head being married and having attained a secondary or a tertiary education; conversely, the proportions of female heads

Table 4 Average characteristics of households in each latent class

	Class 1	Class 2	Class 3
Class size	517	1709	931
Proportion	0.164	0.541	0.295
Time-varying covariates			
disposable income (×1000)	46.061	24.957	37.069
children	0.171	0.161	0.218
married	0.741	0.613	0.710
household size (number)	2.569	2.515	2.680
employee	0.534	0.356	0.560
retired	0.432	0.540	0.391
debts	0.215	0.164	0.224
real assets (×1000)	391.660	163.360	264.309
unemployment rate (%)	6.383	10.553	7.462
bank branch density	6.540	5.055	6.134
Time-constant covariates			
north	0.710	0.335	0.604
centre	0.180	0.177	0.215
small town	0.251	0.300	0.328
female	0.201	0.401	0.282
lower secondary education	0.271	0.348	0.354
upper secondary education	0.410	0.153	0.371
tertiary education	0.259	0.054	0.125
youth stock return (%)	8.804	10.577	11.169

and of those with a lower secondary education show a strong decreasing tendency. Furthermore, households allocated to Class 1 mainly live in regions characterized by a lower average unemployment rate and a higher value of bank branch density, as opposite to Class 2 that presents the highest value of the average unemployment rate and the smallest value of the bank branch density. Class 3 shows characteristics that are intermediate with respect to the first two classes.

4.4 Life-cycle and time patterns of households' investment behavior

Table 5 shows the estimates of coefficients $\boldsymbol{\beta}_u$ and $\boldsymbol{\gamma}_u$ (and the corresponding standard errors) of the fourth-order age and time polynomials for the participation and outcome equations, respectively. The corresponding class-specific age and time profiles (together with 95% confidence bands) are plotted in Fig. 3. These trajectories are estimated considering an individual with mean or modal characteristics (in the case of quantitative and qualitative covariates, respectively).

Focusing on the life-cycle pattern of the probability of participating in risky financial markets (Fig. 3, left graph of panel (a)), we notice a significant heterogeneity

Table 5 Estimated coefficients of age and year polynomials ($\boldsymbol{\beta}_u$ and $\boldsymbol{\gamma}_u$, $u = 1, 2, 3$)

	(1) Participation probability			(2) Share invested		
	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	$\hat{\boldsymbol{\gamma}}_1$	$\hat{\boldsymbol{\gamma}}_2$	$\hat{\boldsymbol{\gamma}}_3$
age	0.062	−0.059	−0.132$^{(.)}$	0.036*	−0.065	0.016
	(0.093)	(0.072)	(0.070)	(0.015)	(0.041)	(0.019)
age^2	0.050	−0.243***	−0.039	0.002	0.094**	0.008
	(0.058)	(0.060)	(0.039)	(0.010)	(0.033)	(0.012)
age^3	−0.006	0.023	0.004	−0.001	0.005	0.002
	(0.017)	(0.015)	(0.015)	(0.003)	(0.008)	(0.005)
age^4	−0.009	0.030***	−0.007	−0.002	−0.012*	0.000
	(0.009)	(0.008)	(0.007)	(0.002)	(0.005)	(0.002)
year	−0.044	−0.006	−0.148***	−0.011$^{(.)}$	0.097***	−0.039***
	(0.038)	(0.035)	(0.025)	(0.006)	(0.022)	(0.006)
year^2	0.341***	0.720***	1.029***	0.069$^{(.)}$	0.119	−0.039
	(0.232)	(0.221)	(0.144)	(0.037)	(0.112)	(0.044)
year^3	0.223	0.098***	0.704	0.065	−0.503**	0.116*
	(0.274)	(0.271)	(0.176)	(0.048)	(0.168)	(0.051)
year^4	−1.373	−2.374*	−5.101***	−0.540*	−0.711	0.271
	(1.333)	(1.251)	(0.842)	(0.226)	(0.680)	(0.258)

Standard errors in round brackets
***p value ≤ 0.001; **p value ≤ 0.01; *p value ≤ 0.05; $^{(.)}$ p value ≤ 0.10

across latent classes. Households in Class 1 are characterized by the highest asset market participation rates (around 70%), whereas households in Class 2 have a very low propensity to invest in risky assets (lower than 3%); in both cases the estimated coefficients of the trajectories are substantially constant over the life-cycle. This is a somewhat expected result and is consistent with the existence of fixed entry costs. These two latent classes are, in fact, characterized by the highest and lowest economic conditions, in terms of average disposable income and real asset wealth, respectively (see Table 4 and related comments). As discussed in Guiso et al. (2003a), in the presence of fixed participation costs only relatively wealthier investors enter risky financial markets, while poor households do not hold risky assets, because the utility loss from abstaining from participation is too small to offset entry costs. The figure also documents a distinct hump-shaped age pattern of participation probability for Class 3: asset market participation increases over the first part of the life-cycle, peaking at the age of approximately 42, then it gradually decreases until the age of 65, whereas the drop is much steeper after retirement. At its peak, the participation rate of Class 3 is around 37%, while at early and later stages of the life-cycle only a small fraction of households invest in risky assets (around 22% and 7%, respectively). This estimated age profile is in line with the findings of Guiso and Jappelli (2002) for Italy and is consistent with the hump-shaped life-cycle patterns estimated in several countries, as found by Guiso et al. (2002) and Guiso et al. (2003b).

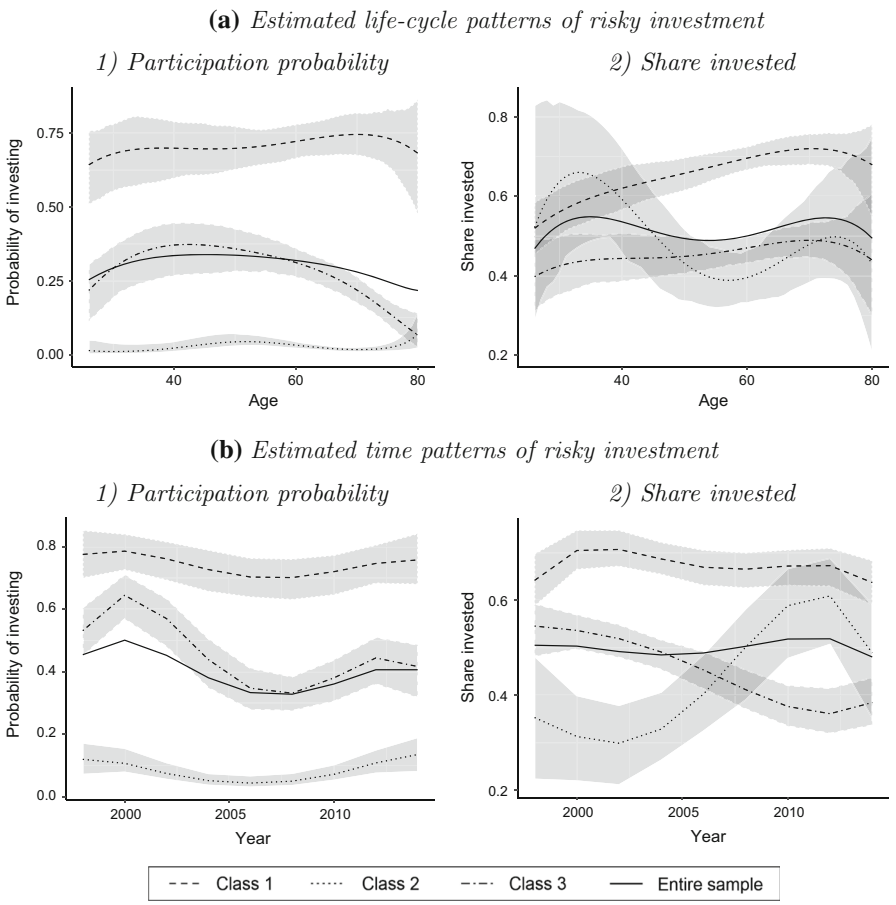

Fig. 3 Estimated age (**a**) and year (**b**) latent trajectories

The average age profile for the entire sample is similar to that of Class 3 and coherent with the theoretical predictions of the life-cycle model and with the empirical findings of the prevailing literature, confirming the still limited asset market participation in Italy.

Regarding the age patterns of the conditional risky assets shares (Fig. 3, right graph of panel (a)), Class 1 and Class 3 are characterized by relatively flat profiles. In particular, households in Class 1 show the highest conditional portfolio shares over most of the life-cycle, reaching the 70% of total financial wealth in later stages. This latent class, composed of households with the highest levels of economic resources and educational attainments, is not only characterized by the highest participation rates, but also by investing more in risky financial assets. This evidence is consistent with the results of previous empirical studies that pointed out the tendency of richer households to specialize in risky financial assets; see Guiso et al. (2002) and Guiso et al. (2003b). Class 2 is instead characterized by a sinusoidal trend along the life-

cycle, with the conditional risky share increasing up to the age of 35, decreasing up to the age of 55, and then slightly increasing again in the last part of the life-cycle. Households in this latent class invest a rather high share in risky asset, especially in the first part of the life-cycle, coherently with theoretical models implying that young households with limited resources should be willing to invest a larger proportion of their wealth in risky financial assets to exploit the higher expected returns of these investments (Haliassos 2003).

The average life-cycle pattern for the entire sample is relatively flat: households maintain the share invested in risky assets fairly constant at around the 55% of their financial wealth and do not engage in substantial rebalancing of their portfolios as they age. This result is in line with the cross-country evidence obtained by Guiso et al. (2003a) and with the findings of the main empirical literature (Ameriks and Zeldes 2004; Alessie et al. 2004).

The estimated time profiles (Fig. 3, panel (b)) confirm the heterogeneity of portfolio choices across latent classes. Participation probability for households in Classes 1 and 2 remains stable over the 1998–2014 period; conversely, a sinusoidal trend is observed for Class 3, with asset market participation decreasing from 2000 to 2008, and increasing in 2010 and 2012. Again, the average profile for the whole sample is similar, but flatter than that of Class 3. Focusing on the time patterns of the conditional share, we find a significant decreasing trend for Class 3, whereas Class 1 is characterized by the highest investment shares, which remain substantially constant over the whole period. A significant sinusoidal trend is estimated for Class 2, with conditional shares decreasing from 1998 to 2002, increasing up to 2012, and then decreasing again in 2014. The average profile is completely flat, with a conditional share constant over the whole period at around 53%. Household portfolio choices in Italy are thus rather stable over time. Business cycle and changing market conditions mainly affect participation probability, which slightly reduces over time. Furthermore, the global financial crisis seems to have had a limited impact on household decision to enter/exit the risky financial market and on portfolio rebalancing. Our results are consistent with the findings of Brunnermeier and Nagel (2008), Calvet et al. (2009), and Bilias et al. (2010), who show that households do not frequently adjust their portfolios and that portfolio rebalancing is not strongly affected by market fluctuations.

4.5 Effect of time-varying and time-constant covariates

The estimated regression coefficients (and the corresponding standard errors) of the remaining time-varying covariates are reported in Table 6. Since the effects of the covariates are allowed to be class-specific, in most cases the statistical significance and the direction of the effect (positive or negative) may change from one class to another. The first column of the table shows the estimated coefficients for the participation equation. Disposable income and real asset wealth exert positive and statistically significant effects on market participation in all the three classes, confirming the crucial role of household economic conditions on the decision of whether to enter risky asset markets. Household size exerts heterogenous effects on market participation: it significantly increases the probability of investing in risky assets for households in Class 3,

in line with the findings of Guiso and Jappelli (2002) and Alessie et al. (2004), whereas it reduces participation probability for Class 2. It is also worth remarking that all the three considered identification variables exert significant effects on the participation probability of all classes, supporting the validity of our identification strategy.

Turning to the conditional investment share (second column of Table 6), we again point out significant heterogeneity in the effects of time-varying covariates. In particular, estimated coefficients are statistically significant mainly for households in Class 2: the conditional risky share for this class is significantly lower for larger households with children and for those with lower disposable income and whose head is an employee or is retired.

Average marginal effects, computed as in (9) and reported in Table 7, may help to assess the overall impact of time-varying covariates. As expected, positive and statistically significant marginal effects on market participation probability are found for disposable income and real asset wealth. Similarly, households living in regions with a high bank branch density and those with married head are more likely to invest in risky assets. The marginal effects for all the remaining covariates are not statistically different from zero, as the opposing effects across latent classes tend to balance each other out.

Analyzing the marginal effects on the conditional investment share, we notice that only the presence of children under 14 year and the occupational status of the household head significantly affect the conditional share invested. Conversely, household disposable income, despite having a substantial influence on market participation, does not exert any significant impact on portfolio allocation.

The estimates of coefficients $\boldsymbol{\delta}_u$ ($u = 2, 3$) of time-constant covariates in the multinomial logit submodel of latent class membership are reported in Table 8, together with the related standard errors. Households living in the Centre and in the North of Italy and the head of which is a male, with a lower or upper secondary and, to a greater extent, a tertiary education, have a lower probability of belonging to Classes 2 and 3 than to Class 1.

Average marginal effects, computed as in (10) and reported in Table 9, allow us to assess the indirect impact of time-constant covariates on both asset market participation and conditional share invested. Female-headed households have a 5.6% lower participation probability, while households living in the Centre and in the North of Italy are 10.7% and 16.7% more likely to invest in risky assets, respectively. Furthermore, the probability of participating to risky asset markets for households whose head has a lower secondary, an upper secondary and a tertiary education is 10.2%, 20.2%, and 25.0% higher than those with no or primary education, respectively. This evidence supports the hypothesis of information-related barriers to asset market participation. Coherently with the findings of most empirical studies (see Guiso et al. 2003b), better-educated households are more likely to invest in risky assets because they are better informed about the existence and properties of different assets, and they are thus more able to take advantage of investment opportunities (Guiso et al. 2003a).

The marginal effects on the share invested are rather small and statistically not significant. However, the conditional risky share is significantly higher for households whose head has a tertiary education, confirming the key role played by educational attainments on household risky financial investment decisions.

Table 6 Estimated coefficients of time-varying covariates (β_u and γ_u, $u = 1, 2, 3$)

	(1) Participation probability			(2) Share invested		
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
disposable income	0.088***	0.098***	0.233***	0.003	−0.014**	0.000
	(0.024)	(0.015)	(0.022)	(0.002)	(0.005)	(0.003)
children	0.187	0.064	−0.082	−0.002	−0.118*	−0.002
	(0.135)	(0.107)	(0.076)	(0.022)	(0.057)	(0.021)
married	−0.028	0.014	0.228***	0.053**	−0.087$^{(.)}$	0.018
	(0.113)	(0.092)	(0.071)	(0.019)	(0.050)	(0.021)
household size	0.041	0.102**	−0.136***	−0.023**	−0.032$^{(.)}$	-0.008
	(0.054)	(0.039)	(0.032)	(0.008)	(0.018)	(0.008)
employee	−0.198	0.226	0.112	0.038	−0.256***	−0.024
	(0.182)	(0.145)	(0.115)	(0.034)	(0.078)	(0.036)
retired	−0.161	0.127	0.041	−0.002	−0.284***	0.083*
	(0.207)	(0.152)	(0.131)	(0.037)	(0.085)	(0.039)
debts	−0.123	0.074	−0.030	0.016	0.077*	−0.024
	(0.088)	(0.083)	(0.058)	(0.016)	(0.040)	(0.016)
real assets	0.001	0.003***	0.004***			
	(0.001)	(0.001)	(0.001)			
unemployment rate	−0.033$^{(.)}$	−0.038*	0.028*			
	(0.020)	(0.016)	(0.012)			
bank branch density	0.133*	0.046	0.178***			
	(0.055)	(0.038)	(0.032)			
constant	−0.165	−2.412***	−2.429***	0.679***	0.962***	0.399***
	(0.497)	(0.362)	(0.305)	(0.041)	(0.114)	(0.053)

Standard errors in round brackets

***p value ≤ 0.001; **p value ≤ 0.01; *p value ≤ 0.05; $^{(.)}$$p$ value ≤ 0.10

Table 7 Marginal effects of time-varying covariates on the probability of participating and on the share invested

	(1) Participation probability	(2) Share invested
disposable income	0.003***	−0.006
	(0.000)	(0.006)
children	0.005	−0.060***
	(0.014)	(0.006)
married	0.023$^{(.)}$	−0.028
	(0.015)	(0.033)
household size	−0.006	−0.023
	(0.006)	(0.015)
employee	0.016	−0.129*
	(0.021)	(0.060)
retired	0.004	−0.114$^{(.)}$
	(0.022)	(0.065)
debts	−0.005	0.033
	(0.010)	(0.031)
real assets	0.001**	
	(0.000)	
unemployment rate	−0.001	
	(0.003)	
bank branch density	0.028***	
	(0.006)	

Bootstrap (199 replications) standard errors in round brackets
***p value ≤ 0.001; **p value ≤ 0.01; *p value ≤ 0.05; $^{(.)}$ p value ≤ 0.10

Finally, as in Fagereng et al. (2017), cohort effects captured by stock market returns experienced in youth have a positive effect on participation probability and a negative effect on the share invested, even if both effects are rather small.

5 Conclusions

In this paper, we propose a bivariate latent growth model to explain longitudinal data when the observation of a response variable of interest is conditioned on a selection mechanism. In particular, we introduce a selection model component with two variables: a binary one that drives the selection phase, and a continuous one, which represents the outcome of main interest. We also rely on a discrete latent variable, which defines unobservable clusters so as to account for different behaviors in the population, defined in terms of latent trajectories.

For estimating the proposed model, we develop an EM algorithm that also relies on an acceleration step based on a suitable numerical algorithm. The computation of standard errors for model parameters, the choice of the number of latent classes

Table 8 Estimated coefficients of time-constant covariates ($\boldsymbol{\delta}_u$, $u = 2, 3$)

	(1) Class 2 membership $\hat{\boldsymbol{\delta}}_2$	(2) Class 3 membership $\hat{\boldsymbol{\delta}}_3$
north	−2.420***	−0.905***
	(0.239)	(0.246)
centre	−1.609***	−0.514$^{(.)}$
	(0.256)	(0.272)
small town	0.085	0.203
	(0.154)	(0.155)
female	0.924***	0.556***
	(0.170)	(0.171)
lower secondary education	−1.596***	−0.652*
	(0.253)	(0.293)
upper secondary education	−2.971***	−1.198***
	(0.269)	(0.296)
tertiary education	−3.534***	−1.810***
	(0.297)	(0.322)
youth stock return	−0.003	0.013*
	(0.007)	(0.007)
constant	4.345***	2.018***
	(0.333)	(0.382)

Class 1 is used as the reference group. Standard errors in round brackets
***p value ≤ 0.001; **p value ≤ 0.01; *p value ≤ 0.05; $^{(.)}$ p value ≤ 0.10

(unobservable clusters), and the clustering of the sample units based on the posterior probabilities of the latent variable are also dealt with.

The proposed approach is motivated by an application on household portfolio choices in Italy over the 1998–2014 period, in terms of both asset market participation and the conditional share invested in risky assets.

Differently from the prevalent literature, which ignores the heterogeneity in household investment choices, we are able to provide an explanation to the empirical inconsistencies observed in previous studies, by clustering households in a finite number of latent classes characterized by heterogeneous investment behaviors over the life-cycle and over time. Specifically, we identify a latent class of households (which represents about 30% of the sample) whose behavior in terms of risky asset market participation follows a hump-shaped trend along the life-cycle. This is consistent with the hump shape in the labor income process and with the existence of significant fixed participation costs in earlier and later stages of the life-cycle. At the same time, we also find that more than one half of the households in the sample do not participate to the risky asset market, confirming a well-established stylized fact in the household portfolio literature. Conversely, the remaining 16% of the households are characterized by a high propensity to invest along all their life-cycle. As far as the share invested in risky financial assets is concerned, we find that the conditional portfolio share for

Table 9 Marginal effects of time-constant covariates on the probability of participating and on the share invested

	(1) Participation probability	(2) Share invested
north	0.167***	− 0.023
	(0.019)	(0.041)
centre	0.107***	− 0.013
	(0.016)	(0.030)
small town	− 0.002	− 0.007
	(0.008)	(0.008)
female	− 0.056***	− 0.005
	(0.010)	(0.011)
lower secondary education	0.102***	− 0.009
	(0.010)	(0.023)
upper secondary education	0.202***	− 0.012
	(0.014)	(0.040)
tertiary education	0.250***	0.024***
	(0.020)	(0.004)
youth stock return	0.000	$-0.001^{(.)}$
	(0.000)	(0.000)

Bootstrap (199 replications) standard errors in round brackets
***p value ≤ 0.001; **p value ≤ 0.01; *p value ≤ 0.05; $^{(.)}p$ value ≤ 0.10

the entire sample remains fairly constant over the life-cycle. In particular, households with an hump-shaped age profile of market participation show a substantially flat trend in the share invested, while those with a high propensity to invest in risky assets are characterized by a slightly increasing trend over the life-cycle.

Our empirical findings suggest that household portfolio choices over the life-cycle mainly concern the decision to enter and exit the market for risky assets, whereas the rebalancing portfolio composition has limited relevance. Moreover, heterogeneity in asset market participation patterns is deeply related to the differences in economic conditions, exposure to background risk, and attitudes towards risk that characterize households belonging to the different latent classes and observed at different stages of their life-cycle.

Acknowledgements S. Bacci acknowledges the financial support provided by the "Dipartimenti Eccellenti 2018-2022" Italian ministerial funds. F. Bartolucci acknowledges the financial support from the grant "Partial effects in econometric models for binary longitudinal data based on quadratic exponential distributions" of the University of Perugia (RICBASE2018).

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement.

Appendix A: simulation study

We illustrate the results of a Monte Carlo simulation study aimed at comparing, in terms of computing time, the performance of the proposed EM algorithm with acceleration with respect to the conventional EM algorithm.

Design

We randomly drew 100 samples from the proposed model with a sample size equal to $n = 1000$ and $T_i = 5$ time occasions for all i. The simulation design assumes the existence of time-varying covariates affecting the distribution of the two response variables, which are randomly generated from a standard Gaussian distribution. Moreover, in both vectors $\boldsymbol{w}_{it}$ and $\boldsymbol{x}_{it}$, a polynomial of order $r = 1$ is included for the year of interview. Two time-constant covariates are assumed to affect the distribution of the latent variable.

We considered a number of latent classes (k) equal to 3 and 4 and two different specifications of the model parameters as follows:

- Scenario 1: $k = 3$, $\rho = -0.5$, $\sigma^2 = 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} 2 & 0 & -2 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \\ 0 & -0.2 & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \\ 0.1 & -0.1 & 0 \end{bmatrix}, \quad \boldsymbol{\Delta} = \begin{bmatrix} 0 & 0 \\ 0.5 & 1 \\ -0.5 & -1 \end{bmatrix},$$

  where, for all matrices the first line is referred to the intercept term.
- Scenario 2: $k = 3$, $\rho = -0.5$, $\sigma^2 = 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} 0 & 0.2 & 0.4 \\ 1 & 0.6 & 0.9 \\ -1 & -1 & -1 \\ 0 & -0.2 & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} 0 & 0.5 & 0.3 \\ -1 & -0.6 & -0.9 \\ 1 & 1 & 1 \\ 0.1 & -0.1 & 0 \end{bmatrix}, \quad \boldsymbol{\Delta} = \begin{bmatrix} 0 & 0 \\ 0.5 & 1 \\ -0.5 & -1 \end{bmatrix}.$$

- Scenario 3: $k = 4$, $\rho = -0.5$, $\sigma^2 = 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} 2 & 0 & -2 & 1 \\ -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 0 & -0.2 & 0 & 0.1 \end{bmatrix}, \boldsymbol{\Gamma} = \begin{bmatrix} -1 & 0 & 1 & 0.5 \\ -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 0.1 & -0.1 & 0 & 0.5 \end{bmatrix}, \boldsymbol{\Delta} = \begin{bmatrix} 0 & 0 & 0 \\ 0.25 & 0.5 & 1 \\ -0.25 & -0.5 & -1 \end{bmatrix}.$$

For all scenarios, in order to estimate model parameters, we run the conventional EM algorithm and the proposed EM with acceleration step on the basis of different tolerance levels ($\epsilon^* = 0.1, 0.01, 0.001, 0.0001$) for switching from the EM steps to the quasi-Newton steps.

Results

In this simulation study we are interested in the computational costs of the algorithms under comparison. However, it is important to underline that all algorithms have reached the convergence at the same maximum of the model log-likelihood. Moreover, to perform a fair comparison, these algorithms have been implemented in R and run on the same personal computer. Table 10 shows the ratio between the average computing time, over the simulated samples, of the conventional EM algorithm and the proposed approach based on the different tolerance levels, under the three scenarios. Table 11 also reports the average number of EM iterations required by the algorithms under comparison to reach the convergence.

From the results, we observe that the proposed acceleration step allows us to achieve the convergence with a lower computational cost with respect to the EM without acceleration. The gain in terms of computing time is more evident when the tolerance level ϵ^* is higher and under the most complex scenarios with regard to parameter estimation and number of latent classes. In any case, even under the worst scenario, the highest computing time is, in average, of the order of some minutes. Moreover, since the proposed EM algorithm relies on an acceleration step based on quasi-Newton methods, it is able to reach the convergence with a lower number of EM steps. The average number of EM iterations increases when $k = 4$ and under Scenario 2, which assumes a more complex structure of model parameters.

Table 10 Ratio between the average computing time of the classical EM algorithm and the proposed EM with acceleration, based on different levels of ϵ^*

	ϵ^*			
	0.1	0.01	0.001	0.0001
Scenario 1	3.245	2.073	1.622	1.167
Scenario 2	7.559	7.557	6.326	2.770
Scenario 3	3.851	2.145	1.629	1.199

Table 11 Average number of EM iterations required by the proposed EM algorithm with acceleration, based on different levels of ϵ^*, and by the classical EM algorithm to reach the convergence

	ϵ^*				
	0.1	0.01	0.001	0.0001	Classical EM
Scenario 1	3.00	5.89	8.02	12.57	33.01
Scenario 2	7.34	7.34	6.77	18.28	308.53
Scenario 3	3.03	7.57	11.21	16.75	40.95

Appendix B: R codes and functions

In the following we provide an example of R script to estimate a bivariate latent growth model as the one proposed in the paper. The dataset we use is available, together with all the estimation functions, at the web page https://github.com/Silvia-Pand/BivLT; it mimics, in a simplified way, the general structure of the data used in the paper. In more detail, the example dataset consists of 1,000 individuals followed up to 9 time occasions. We include 2 (continuous) time-varying covariates and 3 (two binary and one continuous) time-constant covariates. A polynomial of order 2 for a continuous time-varying variable is added to account for the non-linear time effect. We also assume $k = 3$ latent classes.

The script below starts with the preparation of data (arrays of response variables and covariates) and the estimation of the model. Then, the main output corresponding to estimated class weights and regression coefficients of covariates is displayed.

```
> rm(list=ls())
> source("est_biv_LT.R") # estimation function
> source("lk_sel_comp.R")
> source("sc_sel_comp.R")
> source("lk_sel.R")
> source("sc_sel.R")
>
> load("ExampleData.RData") # load data
>
> ####
>
> ## Prepare data
>
> # Response variables
> Y <- Ys # matrix of continuous data (dimension n x TT)
> dim(Y)
[1] 1000    9
> head(Y)
[,1]     [,2] [,3]      [,4] [,5] [,6]      [,7] [,8] [,9]
[1,]    0 1.124023    0 0.0000000   NA   NA        NA   NA   NA
```

```
[2,]    0 0.000000    0 0.0000000    0    0 0.4242629    0    0
[3,]    0 0.000000    0 0.0000000    0    0 0.0000000   NA   NA
[4,]    0 0.000000    0 0.3959001    0    0 0.0000000    0   NA
[5,]    0 0.000000    0 0.0000000    0    0 0.0000000    0   NA
[6,]    0 0.000000    0 0.0000000    0    0 0.0000000    0    0

> B <- Bs # matrix of binary data (dimension n x TT)
> dim(B)
[1] 1000    9
> head(B)
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]    0    1    0    0   NA   NA   NA   NA   NA
[2,]    0    0    0    0    0    0    1    0    0
[3,]    0    0    0    0    0    0    0   NA   NA
[4,]    0    0    0    1    0    0    0    0   NA
[5,]    0    0    0    0    0    0    0    0   NA
[6,]    0    0    0    0    0    0    0    0    0

> n <- dim(Y)[1]  # sample size
> TT <- dim(Y)[2] # number of time occasions

> # Covariates

> XX <- XX1 # matrix of time-varying covariates (plus intercept) affecting Y
> dim(XX1) # dimension (n x TT x (ncovX+1))
[1] 1000    9    2

> WW <- WW1 # matrix of time-varying  covariates (plus intercept) affecting B
> dim(WW1) # dimension (n x TT x (ncovW+1))
[1] 1000    9    3

> ZZ <- ZZ1 # matrix of time-constant covariates affecting class membership
> dim(ZZ) # dimension (n x ncovZ)
[1] 1000    3

> ncovX <- dim(XX)[3]-1
> ncovX
[1] 1

> ncovW <- dim(WW)[3]-1
> ncovW
[1] 2

> ncovZ <- dim(ZZ)[2]
> ncovZ
[1] 3

> # include a second-order polynomial for year (year and year^2)
> XXn <- array(0,c(n,TT,ncovX+3))
> XXn[,,1:(ncovX+1)] <- XX
> XXn[,,ncovX+2] <- rep(1,n)%o%((1:TT)-mean(1:TT))
> XXn[,,ncovX+3] <- XXn[,,ncovX+2]^2/10
> dim(XXn)
[1] 1000    9    4
> head(XXn[,1,])
[,1]     [,2] [,3] [,4]
[1,]    1 8.005082   -4  1.6
[2,]    1 2.076157   -4  1.6
[3,]    1 1.962536   -4  1.6
[4,]    1 4.364061   -4  1.6
```

```
[5,]    1 4.229782   -4  1.6
[6,]    1 1.358798   -4  1.6

> WWn <- array(0,c(n,TT,ncovW+3))
> WWn[,,1:(ncovW+1)] <- WW
> WWn[,,ncovW+2] <- rep(1,n)%o%((1:TT)-mean(1:TT))
> WWn[,,ncovW+3] <- WWn[,,ncovW+2]^2/10
> dim(WWn)
[1] 1000    9    5
> head(WWn[,1,])
[,1]     [,2]      [,3] [,4] [,5]
[1,]    1 8.005082 41.523135   -4  1.6
[2,]    1 2.076157 11.103823   -4  1.6
[3,]    1 1.962536  1.394434   -4  1.6
[4,]    1 4.364061 15.596999   -4  1.6
[5,]    1 4.229782 31.297287   -4  1.6
[6,]    1 1.358798 56.965192   -4  1.6

> ####

> ## Model estimation

> k <- 3    # number of latent classes

> # Model estimation with deterministic initialization
> # Warning: it takes a lot of time! Results are stored in ResultsSimData.RData
> out = est_biv_LT(Y, B, XXn, WWn, Z = ZZ,
+ k = k, start = 0,st.err=TRUE)

> # Model estimation with random initialization
> # Warning: it takes a lot of time! Results are stored in ResultsSimData.RData
> set.seed(321)
> outr = list()
> for(i in 1:5) {
+ outr[[i]] = try(est_biv_LT(Y, B, XXn, WWn, Z = ZZ,
+ k = k, start = 1,st.err = TRUE))
}

> save.image("ResultsSimData.RData")

> ####

> ## Display output
> load("ResultsSimData.RData")

> LK = rep(0,6)
> LK[1]  = out$lk
> for(i in 1:5) LK[i+1] = outr[[i]]$lk
> LK
[1] -1029.927 -1029.926 -1052.069 -1029.926 -1029.926 -1029.926

> BIC = rep(0,6)
> BIC[1]  = out$bic
> for(i in 1:5) BIC[i+1] = outr[[i]]$bic
> BIC
[1] 2280.901 2280.900 2325.187 2280.900 2280.900 2280.900

> # Remark: there are not particular problems of local maxima
> # We retain model with deterministic initialization (out)
```

```
> # Averaged weights
> colMeans(out$Piv)
[1] 0.09435604 0.01243584 0.89320812

> out$rho     # correlation between Y and B
[1] -0.5109823
> out$si2
[1] 0.07739475

> ## Effects of covariates

> # effects on B (binary outcome)
> coefGa <- round(out$Ga, 4)
> seGa <- round(out$seGa, 4)
> coefGa # dimension (ncovW x k)
[,1]    [,2]    [,3]
[1,] -2.3671 -0.7252 -2.5248
[2,]  0.1653  0.4212  0.1370
[3,]  0.0064 -0.0446  0.0015
[4,]  0.1054 -0.0553  0.0114
[5,]  0.4386  0.0788  0.0384
> seGa
[,1]   [,2]   [,3]
[1,] 0.2792 0.4081 0.0853
[2,] 0.0473 0.1874 0.0155
[3,] 0.0029 0.0164 0.0007
[4,] 0.0416 0.0732 0.0171
[5,] 0.1709 0.2862 0.0696

> # effects on Y (continuous outcome)
> coefBe = round(out$Be, 4)
> seBe = round(out$seBe, 4)
> coefBe # dimension (ncovX x k)
[,1]    [,2]    [,3]
[1,]  0.5698  1.0118  1.2211
[2,] -0.0070 -0.0467 -0.0230
[3,]  0.0075  0.0284  0.0006
[4,] -0.0107 -0.0396  0.0360
> seBe
[,1]   [,2]   [,3]
[1,] 0.1636 0.1276 0.1667
[2,] 0.0101 0.0291 0.0063
[3,] 0.0153 0.0204 0.0095
[4,] 0.0648 0.0828 0.0422

> # effects on latent classes
> coefDe = round(out$De, 3)
> seDe = round(out$seDe, 3)
> coefDe # dimension (ncovZ x (k-1)); reference: Class 1
[,1]   [,2]
[1,] -2.263  2.355
[2,]  0.336  0.006
[3,] -9.681  0.940
[4,]  0.019 -0.025
> seDe
[,1]  [,2]
[1,] 0.730 0.416
[2,] 0.861 0.474
[3,] 0.000 0.832
[4,] 0.034 0.018
```

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second international symposium of information theory. Akademiai Kiado, Budapest, pp 267–281

Alessie R, Hochguertel S, van Soest A (2004) Ownership of stocks and mutual funds: a panel data analysis. Rev Econ Stud 86:783–796

Ameriks J, Zeldes SP (2004) How do household portfolio shares vary with age? Working paper, Columbia University

Atella V, Brunetti M, Maestas N (2012) Household portfolio choices, health status and health care systems: a cross-country analysis based on share. J Bank Finance 36:1320–1335

Bacci S, Pandolfi S, Pennoni F (2014) A comparison of some criteria for states selection in the latent Markov model for longitudinal data. Adv Data Anal Classif 8:125–145

Bacci S, Bartolucci F, Bettin G, Pigini C (2019) A latent class growth model for migrants' remittances: an application to the German socio-economic panel. J R Stat Soc Ser A 182:1607–1632

Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ (1997) Latent variable regression for multiple discrete outcomes. J Am Stat Assoc 92:1375–1386

Bank of Italy (2015) Household Income and Wealth in 2014, Supplement to the Statistical Bulletin No. 64, Bank of Italy, Rome

Bartolucci F, Farcomeni A (2009) A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. J Am Stat Assoc 104:816–831

Bartolucci F, Murphy B (2015) Finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. J Quant Anal Sports 11:193–203

Bilias Y, Georgarakos D, Haliassos M (2010) Portfolio inertia and stock market fluctuation. J Money Credit Bank 42:715–742

Bollen KA, Curran PJ (2006) Latent curve models: a structural equation perspective. Wiley, Hoboken

Brandolini A (1999) The distribution of personal income in post-war Italy: source description, data quality, and the time pattern of income inequality. Giornale degli Economisti 58:183–239

Bray BC, Lanza ST, Tan X (2015) Eliminating bias in classify-analyze approaches for latent class analysis. Struct Equ Model 22:1–11

Brunnermeier MK, Nagel S (2008) Do wealth fluctuations generate time-varying risk aversion? Micro-evidence on individuals. Am Econ Rev 98:713–736

Calvet LE, Campbell JY, Sodini P (2009) Fight or flight? Portfolio rebalancing by individual investors. Q J Econ 124:301–348

Christelis D, Jappelli T, Padula M (2010) Cognitive abilities and portfolio choice. Eur Econ Rev 54:18–38

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 39:1–38

Dias JG, Vermunt JK (2008) A bootstrap-based aggregate classifier for model-based clustering. Comput Stat 23:643–659

Edwards RD (2008) Health risk and portfolio choice. J Bus Econ Stat 26:472–485

Efron B, Tibshirani R (1993) An Introduction to the Bootstrap. Chapman & Hall, New York

Fagereng A, Gottlieb C, Guiso L (2017) Asset market participation and portfolio choice over the life-cycle. J Finance 72:705–750

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631

Giuliano P, Spilimbergo A (2013) Growing up in a recession. Rev Econ Stud 81:787–817

Givens GH, Hoeting JA (2013) Computational statistics. Wiley, Hoboken

Goodman LA (1974) The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A modified latent structure approach. Am J Sociol 79:1179–1259

Goodman LA (2007) On the assignment of individuals to latent classes. Sociol Methodol 37:1–22

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman and Hall, Boca Raton

Guiso L, Jappelli T (2002) Household portfolios in Italy. In: Guiso L, Haliassos M, Jappelli T (eds) Household portfolios, chapter 7. The MIT Press, Cambridge, pp 251–289

Guiso L, Paiella M (2008) Risk aversion, wealth, and background risk. J Eur Econ Assoc 6:1109–1150

Guiso L, Sodini P (2013) Household finance: an emerging field. In: Constantinides G, Harris M, Stulz RM (eds) Handbook of the economics of finance, chapter 21, vol 2. Elsevier, Amsterdam, pp 1397–1532
Guiso L, Haliassos M, Jappelli T (2002) Household portfolios. The MIT Press, Cambridge
Guiso L, Haliassos M, Jappelli T (2003a) Household stockholding in Europe: where do we stand and where do we go? Econ Policy 18:125–170
Guiso L, Haliassos M, Jappelli T (2003b) Stockholding in Europe portfolios. Palgrave Macmillan, New York
Haliassos M (2003) Stockholding: recent lessons from theory and computations. In: Guiso L, Haliassos M, Jappelli T (eds) Stockholding in Europe portfolios. The MIT Press, New York, pp 30–51
Haliassos M (2008) Household portfolios. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, New York, pp 1110–1129
Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47:153–161
Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley series in probability and statistics, 2nd edn. Wiley, New York
Marchenko YV, Genton MG (2012) A Heckman selection-t model. J Am Stat Assoc 107:304–317
McLachlan G, Peel D (2000) Finite mixture models. Wiley, Hoboken
Merton R (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. Rev Econ Stat 51:247–257
Muthén BO (2004) Latent variable analysis: growth mixture modelling and related techniques for longitudinal data. In: Kaplan D (ed) Handbook of quantitative methodology for the social sciences. Sage, Newbury Park, pp 345–368
Muthén BO, Shedden K (1999) Finite mixture modelling with mixture outcomes using the EM algorithm. Biometrics 55:463–469
Nylund KL, Asparouhov T, Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Equ Model 14:535–569
Nylund-Gibson K, Masyn KE (2016) Covariates and mixture modeling: results of a simulation study exploring the impact of misspecified effects on class enumeration. Struct Equ Model 23:782–797
Oakes D (1999) Direct calculation of the information matrix via the EM algorithm. J R Stat Soc B 61:479–482
Puhani PA (2000) The Heckman correction for sample selection and its critique. J Econ Surv 14:53–68
Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
van Rooij M, Lusardi A, Alessie R (2011) Financial literacy and stock market participation. J Financ Econ 101:449–472
Vermunt JK (2010) Latent class modeling with covariates: two improved three-step approach. Political Anal 18:450–469