Causal latent Markov model for the comparison of multiple treatments in observational longitudinal studies

Bartolucci, Francesco and Pennoni, Fulvia and Vittadini, Giorgio

University of Perugia, University of Milano-Bicocca, University of Milano-Bicocca

August 2015

Causal latent Markov model for the comparison of multiple treatments in observational longitudinal studies

Francesco Bartolucci
Department of Economics
University of Perugia (IT)
email: francesco.bartolucci@unipg.it

Fulvia Pennoni
Department of Statistics and Quantitative Methods
University of Milano-Bicocca (IT)
e-mail: fulvia.pennoni@unimib.it

Giorgio Vittadini
Department of Statistics and Quantitative Methods
University of Milano-Bicocca (IT)
e-mail: giorgio.vittadini@unimib.it

September 7, 2015

Abstract

We extend to the longitudinal setting a latent class approach that has beed recently introduced by Lanza et al. (2013) to estimate the causal effect of a treatment. The proposed approach permits the evaluation of the effect of multiple treatments on subpopulations of individuals from a dynamic perspective, as it relies on a Latent Markov (LM) model that is estimated taking into account propensity score weights based on individual pre-treatment covariates. These weights are involved in the expression of the likelihood function of the LM model and allow us to balance the groups receiving different treatments. This likelihood function is maximized through a modified version of the traditional expectation-maximization algorithm, while standard errors for the parameter estimates are obtained by a non-parametric bootstrap method. We study in detail the asymptotic properties of the causal effect estimator based on the maximization of this likelihood function and we illustrate its finite sample properties through a series of simulations showing that the estimator has the expected behavior. As an illustration, we consider an application aimed at assessing the relative effectiveness of certain degree programs on the basis of three ordinal response variables when the work path of a graduate is considered as the manifestation of his/her human capital level across time.

Keywords: Causal inference, Expectation-Maximization algorithm, Hidden Markov models, Multiple treatments, Policy evaluation, Propensity score.

1 Introduction

We propose a causal inference approach, based on a Latent Markov (LM) model, to dynamically evaluate the average effect of multiple treatments in a longitudinal context in which multivariate responses are observed at different time occasions. For this aim, we formulate the LM model (Bartolucci et al., 2013) within a Potential Outcome (PO) conceptual approach (Rubin, 1974, 2005), and we propose to estimate this model using a Propensity Score (PS) method (Imbens, 2000). An estimator of causal effects results, which is based on a weighted likelihood function, with weights computed according to the PS approach; this estimator may be used in order to evaluate the efficacy of a treatment versus another one on different subpopulations across time.

In experimental settings, techniques for estimating causal effects with more than two treatments have been used for many years. This issue is also of great interest in non-experimental settings, where the need arises of estimating causal effects on the basis of observational data. In this regard, the PO framework is of particular interest as it allows us to define causal effects in a straightforward way and to formulate techniques for their estimation. In its original formulation, a set of POs corresponding to the possible treatments is assumed to exist for every sample unit, but it is possible to observe only the outcome corresponding to the taken treatment. This "missingness" problem is the primary challenge of causal inference and, therefore, suitable methods have been proposed in order to overcome it and then obtain reliable estimators of the causal effects defined in terms of summary statistics of the individual differences between POs. Among these measures, the Average Treatment Effect (ATE) is of particular relevance. It is defined as the average, over the population under study, of the difference between the POs corresponding to two treatments of interest or between the treatment of interest and a control. As other causal effects, ATE may be estimated by a PS method (see, among others, Rosenbaum and Rubin, 1983; Imbens, 2000; Guo and Fraser, 2010), so as to balance the groups corresponding to the different treatments, taking into account pre-treatment covariates. In practice, the estimate of the effect of interest, such as ATE, is computed by using matching methods, as in (Angrist, 1991), or by weighting estimators similar to those proposed by Robins et al. (2000). As noted by many authors, PS weighted estimators follow the well-known sampling theory approach of Horvitz and Thompson (1952) for estimating a population total. Some generalizations have been proposed by Hirano et al. (2003); see also Hernán et al. (2001). It also worth noting that McCaffrey et al. (2013) recently introduced an innovative use of PS weights to control for pre-treatment unbalances on the basis of observed variables in non-randomized/observational studies. The authors suggested an operative pathway to implement PS weighting for multiple treatments using generalized boosted models.

Some authors highlighted the connection between model based causal statistical methods in the tradi-

tional sense and methods of causal inference based on the PO approach. Along these lines, Lanza et al. (2013) proposed a PS based method for making causal inference in connection with the Latent Class (LC) model (Lazarsfeld and Henry, 1968; Goodman, 1974). This approach considers both matching and weighting on the basis of individual PSs depending on pre-treatment covariates. In this paper, we extend the proposal of Lanza et al. (2013) to a longitudinal context in which different measurement occasions are considered after the treatment. In this regard it is worth considering that, in economic and social contexts, we can easily dispose of longitudinal observational datasets and often the focus is on the study of the effect of a policy or treatment on certain outcomes of interest, accounting also for the evolution of such an effect across time on different subpopulations of interest. In the estimation of causal effects, longitudinal data are generally preferable to cross-sectional data; see, among others, Aalen et al. (2012) and Arjas (2013) and the winter 2014 issue of *Econometric Theory*, which gives a comprehensive discussion about methods of causal inference.

As already mentioned, the approach we propose is based on the LM model, which is an important statistical model that may be conceived as an extended version of the LC model having a longitudinal dimension. The LM model dates back to Wiggins (1955), who introduced it to analyze latent changes on the basis of panel data. This initial version is based on a homogeneous Markov chain and may be used with a single outcome at each time occasion. Wiggins (1955) formulated the model so that the manifest transition is a mixture of the true change and a spurious change due to measurement errors in the observed states. Among the main extensions of the basic version of the LM model we mention: those based on the inclusion of constraints on the measurement component or on the latent structure of the model (Bartolucci, 2006); those based on the inclusion of individual covariates (Vermunt et al., 1999; Bartolucci and Farcomeni, 2009); those to deal with multilevel longitudinal data (Bartolucci et al., 2011). For a thorough review see Bartolucci et al. (2013, 2014) and Pennoni (2014).

One of the novelties of the proposal we formulate in this article relies on the adopted parameterization of the latent process, so as to express causal effects in terms of transition between latent states corresponding to subpopulations characterized by different conditions of interest for the study. Moreover, it is also innovative the use of a PS method for the estimation of the causal LM model, which is formulated following a PO approach. In practice, we implement a weighted maximum likelihood estimator based on two steps. At the first step, the probability of choosing a certain treatment on the basis of the pre-treatment covariates is estimated by means of a multinomial logit model. At the second step, a weighted version of the log-likelihood of the LM model, with weights depending on the estimates computed at the first step, is maximized. In order to obtain reliable standard errors for the estimates of the LM parameters, and of the causal effects,

we use a non-parametric bootstrap method (Davison and Hinkley, 1997).

Overall, with multiple treatments and multivariate longitudinal data, the proposed approach allows us to reliably estimate ATEs in terms of initial and transition probabilities between different latent states. In practice, these latent states summarize the different configurations of the outcomes observed at each time occasion. In the proposed formulation the response variables are categorical, but it may be simply generalized to deal with any kind of response variable.

The remainder of this article is organized as follows. The next section outlines the traditional LM model framework. Section 3 provides a technical illustration of our proposal, focusing in particular on the assumptions of the causal LM model and the parametrizations adopted to express causal effects. Section 4 illustrates the two-step estimator of the proposed model based on a PS weighting scheme. Section 5 deals with the consistency of the proposed estimator of the treatment effects and illustrates, under different scenarios, its finite sample properties by means of a simulation study. Section 6 describes an application focused on the evaluation of the effect of different academic degrees based on data collected in the Lombardy region (Italy). In the last section we provide some concluding remarks.

2 Standard latent Markov model

Let n denote the number of individuals that are longitudinally observed and let T denote the number of time occasions. With reference to individual i and occasion t, $i = 1, \ldots, n$, $t = 1, \ldots, T$, we observe a vector of r categorical response variables $\boldsymbol{Y}_{it} = (Y_{i1t}, \ldots, Y_{irt})$. Each response variable Y_{ijt}, $j = 1, \ldots, r$, has c_j categories, labeled from 0 to $c_j - 1$. For every individual i and time occasion t we also consider a vector of covariates $\boldsymbol{X}_{it}$. Throughout this article we denote a realization of random variables or vectors with the lower case.

The variable of main interest in the present approach is a latent variable that is individual- and time-specific; moreover, it has a discrete distribution with support $\{1, \ldots, k\}$ for which more details are provided below. For individual i and time occasion t, this variable is denoted by H_{it}, given its nature of hidden variable, and for the same individual i we introduce the latent vector $\boldsymbol{H}_i = (H_{i1}, \ldots, H_{iT})$. We assume that the variables in $\boldsymbol{Y}_{it}$ are conditionally independent given H_{it} (local independence); we also assume conditional independence between the response variables at different time occasions given the latent process. Note that the assumption that the latent variables are discrete corresponds to defining a certain number k of latent classes (or latent states) of individuals, with individuals in the same class having the same latent characteristics. In practice, these classes correspond to subpopulations having different expected behaviors

with reference to the response variables.

Among the available parametrizations for LM models, in terms of distribution of the response variables and the latent variables, that proposed by Bartolucci et al. (2011) is of particular interest. The resulting model is suitable to deal with educational data having a multilevel structure, due to students collected in classes. Under this model, which is formulated for binary outcomes, the conditional distribution of the response variables given the latent variables is parameterized as

$$p(Y_{ijt} = 1|H_{it} = h) = \frac{\exp(\xi_h - \nu_{jt})}{1 + \exp(\xi_h - \nu_{jt})}, \quad h = 1, \ldots, k,\ j = 1, \ldots, r,\ t = 1, \ldots, T, \tag{1}$$

under the constraint of monotonicity $\xi_1 < \cdots < \xi_k$, where ξ_h is the h-the level of the latent trait characterizing the student, and ν_{jt} is the difficulty level of item j administered at occasion t. The above constraint of monotonicity on the parameters in $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)'$ is typically adopted in the item response theory (see, among others, Hambleton and Swaminathan, 1985). This constraint may be also formulated for the case of ordinal response variables, so that the latent classes (or latent states) corresponding to the different values of H_{it} are suitably ordered (see also Bartolucci et al., 2013).

Under the model of Bartolucci et al. (2011), the initial and transition probabilities of the latent process are parameterized using global logits (McCullagh, 1980). Ignoring for simplicity the multilevel structure, with reference to the initial probabilities we have

$$\log \frac{p(H_{i1} \geqslant h)}{p(H_{i1} < h)} = \alpha_h + \boldsymbol{x}_{i1}'\boldsymbol{\beta}, \quad h = 2, \ldots, k.$$

Concerning the transition probabilities we have

$$\log \frac{p(H_{it} \geqslant h|H_{i,t-1} = \bar{h})}{p(H_{it} < h|H_{i,t-1} = \bar{h})} = \gamma_{\bar{h}ht} + \boldsymbol{x}_{it}'\boldsymbol{\delta}_t,$$

for $\bar{h} = 1, \ldots, k$, $h = 2, \ldots, k$, and $t = 2, \ldots, T$. The intercepts in the above expressions depend on the level of the latent variable and $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_t$ are vectors of regression coefficients for the individual covariates. In this article, we adopt a related parametrization that does not require the constraint of monotonicity so that the conditional distribution of each response variable, given the corresponding latent variable, is free.

3 Causal latent Markov model

For the case of a multiple treatment denoted by the discrete variable Z_i that may depend on a vector of pre-treatment covariates $\boldsymbol{X}_i$, we reformulate the LM from a causal perspective. For this aim, we introduce

"potential versions" of the latent variables H_{it}, which are denoted by $H_{it}^{(z)}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, $z = 1, \ldots, l$, where l is the number of the possible treatments. In particular, $H_{it}^{(z)}$ corresponds to the latent state of individual i at occasion t if he/she had taken treatment z. The sequence of these variables for the same individual i is collected in the vector $\boldsymbol{H}_i^{(z)} = (H_{i1}^{(z)}, \ldots, H_{iT}^{(z)})$ and their distribution may arbitrarily depend on the pre-treatment covariates. As in typical PO approaches, this dependence is not explicitly modeled; more details about this point are provided at the beginning of Section 5.

Variables $H_{it}^{(z)}$ are not directly observable for any z. This is because, as in a typical problem of causal inference, only one of the possible outcomes $H_{it}^{(z)}$ is indeed selected and this selection may depend on the value of the variables themselves. Moreover, in the present context there is a further reason that complicates the inference: the variable of interest is not directly observable even for the selected treatment. This is because it is a latent variable that affects the response vector $\boldsymbol{Y}_{it}$, and then it is only indirectly observable through this vector.

Regarding the latent processes $\boldsymbol{H}_i^{(z)}$, we assume that they follow a first-order Markov chain. The initial probabilities are modeled through a multinomial parametrization depending on the treatment and that is alternative to the parametrization illustrated at the end of the previous section. In particular, regarding the initial and transition probabilities we assume a baseline-category logit model which does not require any ordering of the latent states. Regarding the initial probabilities, we assume that

$$\log \frac{p(H_{i1}^{(z)} = h)}{p(H_{i1}^{(z)} = 1)} = \alpha_h + \boldsymbol{d}(z)'\boldsymbol{\beta}_h, \quad h = 2, \ldots, k, \tag{2}$$

where α_h is the intercept, $\boldsymbol{\beta}_h = (\beta_{h2}, \ldots, \beta_{hl})'$ is a column vector of $l-1$ parameters, and $\boldsymbol{d}(z)$ is a column vector of $l-1$ zeros with the $(z-1)$-th element equal to 1 if $z > 1$. In this way, for $z > 1$, the element β_{hz} of $\boldsymbol{\beta}_h$ corresponds to the effect of the z-th treatment with respect to the first treatment. More precisely, this is an ATE as it is referred to the whole population of interest. For instance, β_{22} being positive means that the second type of treatment increases the probability that the individual is in latent class 2 with respect to first treatment. On the other hand, it is worth noting that, due to the discrete nature of the treatment indicator z, equation (2) does not impose any restriction on the marginal distribution of $H_{i1}^{(z)}$. Moreover, in order to measure ATE for comparing two treatments, we can directly use comparisons of the type $p(H_{i1}^{(z)} = h) - p(H_{i1}^{(1)} = h)$ for $z = 2, \ldots, l$, which are expressed on the probability rather than on the logit scale. Finally, in terms of causal effects, a comparison different from that using the first type of treatment as reference category may be simply based on the difference between suitable elements of $\boldsymbol{\beta}_h$. For instance, provided that $l \geqslant 3$, the third treatment may be compared with the second one through the

difference $\beta_{h3} - \beta_{h2}$ or $p(H_{i1}^{(3)} = h) - p(H_{i1}^{(2)} = h)$, which are again ATEs. Obviously, in an application all these effects will be based on parameter estimates which are computed as described in Section 4.

The transition probabilities of the hidden chain are modeled as follows:

$$\log \frac{p(H_{it}^{(z)} = h | H_{i,t-1}^{(z)} = \bar{h})}{p(H_{it}^{(z)} = 1 | H_{i,t-1}^{(z)} = \bar{h})} = \gamma_{\bar{h}h} + \boldsymbol{d}(z)'\boldsymbol{\delta}_h, \quad \bar{h} = 1, \ldots, k,\ h = 2, \ldots, k,\ t = 2, \ldots, T. \tag{3}$$

The parameters to be estimated in this case are the intercepts $\gamma_{\bar{h}h}$ and the vectors of regression coefficients in $\boldsymbol{\delta}_h = (\delta_{h2}, \ldots, \delta_{hl})'$, $h = 1, \ldots, k$. In particular, each coefficient δ_{hz} is again an ATE which, however, is referred to the transition from level 1 to level h of the latent variable. This effect has an interpretation similar to the one given above in terms of difference on the logit scale between a certain treatment and the first treatment. On the other hand, even in this case we can conceive an ATE directly measured on the probability scale, having expression $p(H_{it}^{(z)} = h | H_{i,t-1}^{(z)} = \bar{h}) - p(H_{it}^{(1)} = h | H_{i,t-1}^{(1)} = \bar{h})$, with $\bar{h} = 1, \ldots, k$, $h = 2, \ldots, k$, or $p(H_{it}^{(z)} = h) - p(H_{it}^{(1)} = h)$, after having properly elaborated the initial and transition probabilities of the Markov chain. Comparisons between two arbitrary treatments in terms of causal effect are also possible and may be based on differences between the corresponding elements of the vector $\boldsymbol{\delta}_h$, such as $\delta_{h3} - \delta_{h2}$, or between transition probabilities, such as $p(H_{it}^{(3)} = h | H_{i,t-1}^{(3)} = \bar{h}) - p(H_{it}^{(2)} = h | H_{i,t-1}^{(2)} = \bar{h})$ with $\bar{h} = 1, \ldots, k$, $h = 2, \ldots, k$.

About the parametrization adopted in the proposed model, we clarify that no constraints are assumed on the distribution of every response variable Y_{ijt} given the corresponding latent variable $H_{it}^{(z)}$. Therefore, this distribution depends on the following parameters:

$$\phi_{jy|h} = p(Y_{ijt} = y | H_{it}^{(z)} = h), \quad h = 1, \ldots, k,\ j = 1, \ldots, r,\ y = 0, \ldots, c_j - 1.$$

The above conditional distribution of the response variables at a specific time occasion, given the corresponding latent state, helps to identify the latent states. Such probabilities may be parametrized as in equation (1), but in the present context we avoid this or similar parametrizations in order to obtain a more flexible model.

The above assumptions incorporate in the model the belief that the only variables that have a direct effect on the outcomes are the latent variables. In particular, local independence means that if the latent variable for an individual at a given occasion was known, the knowledge of an outcome would not help in predicting the other outcomes for the same occasion. Obviously, this does not rule out an effect of the pre-treatment covariates on the response variables, but this effect passes through the latent variables that have,

therefore, a fundamental role. It is also worth noting that the latent variables, as in the model presented in Section 2, are conceived as discrete with k possible values. This allows us to consider a semi-parametric distribution that is parsimonious and based on parameters which are easy to interpret. It is also well known that this assumption allows the model to properly fit the data, as shown in many applications (see also Bartolucci et al., 2013), giving more flexibility to the latent structure with respect to using continuous latent variables. Finally, using discrete latent variables simplifies the computation of the manifest distribution of the response variables given the covariates as it avoids the use of integrals, which are instead necessary with continuously distributed latent variables.

Note that assumptions similar to the previous ones may be formulated with continuous outcomes and also outcomes of a mixed nature can be dealt with. For instance, if some of the outcomes are continuous, these may be modeled by assuming a suitable conditional parametric distribution, such as the Normal distribution, with parameters depending on the latent classes, as in a finite mixture model (McLachlan and Peel, 2000). This makes the proposed approach flexible enough to deal with very different contexts in which longitudinal data are observed and causal effects have to be estimated with reference to the dynamics of the response variables induced by the treatment.

Regarding the treatment selection mechanism, we formulate the following assumptions which are typical of the causal inference literature:

- we assume the consistency rule according to which $H_{it} = H_{it}^{(z_i)}$, where z_i is the observed treatment of individual i;
- we assume that $0 < p(Z_i = z|\boldsymbol{x}_i) < 1$ for $z = 1, \ldots, l$ and any possible configuration $\boldsymbol{x}_i$ of the pre-treatment covariates. In this way, every unit has a positive probability of taking each possible treatment;
- *absence of unmeasured confounding* according to which $Z_i \perp\!\!\!\perp \boldsymbol{H}_i^{(z)}|\boldsymbol{X}_i$, $z = 1, \ldots, l$.

The combination of the last two assumptions is referred to as *strong ignorability* by Rosenbaum and Rubin (1983). This means that, given the pre-treatment covariates, the choice of the treatment is independent of the potential outcomes.

4 Two-step maximum likelihood estimation

In order to estimate the causal model formulated in the previous section and following Lanza et al. (2013), we propose a two-step strategy based on a PS method suitable for multiple treatments (Imbens, 2000;

McCaffrey et al., 2013). At the first step, we estimate a model for the probability of taking a certain type of treatment given the individual pre-treatment covariates. At the second step, we estimate the assumed LM model by maximizing a weighted version of the log-likelihood for this model, with weights based on the parameter estimates from the previous step. It must be clear that the observed data consist, for $i = 1, \ldots, n$, of the vector of pre-treatment covariates $\boldsymbol{x}_i$ and the received treatment z_i, further to the time-specific vectors of responses $\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}$.

In more detail, the first step of the proposed method consists in estimating a multinomial logit model based on the assumption

$$\log \frac{p(Z_i = z|\boldsymbol{x}_i)}{p(Z_i = 1|\boldsymbol{x}_i)} = \eta_z + \boldsymbol{x}_i' \boldsymbol{\lambda}_z, \quad z = 2, \ldots, l, \tag{4}$$

where η_z and $\boldsymbol{\lambda}_z$ are regression parameters. On the basis of the parameter estimates, we compute the individual weights

$$\hat{w}_i = n \frac{1/\hat{p}(z_i|\boldsymbol{x}_i)}{\sum_{m=1}^{n} 1/\hat{p}(z_m|\boldsymbol{x}_i)}, \quad i = 1, \ldots, n. \tag{5}$$

Note that these weights are rescaled so that their sum is equal to the sample size, that is, $\sum_{i=1}^{n} \hat{w}_i = n$; this is necessary for the model selection as we explain below. Moreover, as we illustrate in the application in Section 6.2, the actual set of pre-treatment covariates must be chosen appropriately; see also McCaffrey et al. (2013).

At the second step, we maximize the weighted log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \hat{w}_i \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \log p(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}|z_i), \tag{6}$$

where $\boldsymbol{\theta}$ is the vector of all LM model parameters arranged in a suitable way. The *manifest probability* $p(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}|z_i)$ is computed by suitable recursions developed in the hidden Markov literature (Baum et al., 1970; Welch, 2003; Zucchini and MacDonald, 2009) on the basis of the probabilities $p(\boldsymbol{y}_{it}|H_i^{(t)} = h)$ and the initial and transition probabilities of the corresponding hidden Markov chain parametrized as in equations (2) and (3). Given the assumption of local independence described above, we have that

$$p(\boldsymbol{Y}_{it} = \boldsymbol{y}|H_{it} = h) = \prod_{j=1}^{r} \phi_{jy_j|h},$$

where $\boldsymbol{y} = (y_1, \ldots, y_r)$ is a generic configuration of $\boldsymbol{Y}_{it}$.

The above model log-likelihood may be maximized with respect to $\boldsymbol{\theta}$ by using the Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977), which represents the main tool to estimate dis-

crete latent variable models; see Bartolucci et al. (2013), Chapter 5, for details about its implementation for LM models. The EM algorithm is based on the concept of *complete data*, which correspond to the value of every latent variable, further to the observed covariates and response variables. Therefore, after some algebra, the *complete data log-likelihood* is given by

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_{h=1}^{k}\sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{y=0}^{c_j-1} a_{hjty}\log\phi_{jy|h} + \sum_{h=1}^{k}\sum_{i=1}^{n}\hat{w}_i b_{hi1}\log p(H_{i1}=h|z_i) \\ &\quad + \sum_{\bar{h}=1}^{k}\sum_{h=1}^{k}\sum_{i=1}^{n}\sum_{t=2}^{T}\hat{w}_i b_{\bar{h}hit}\log p(H_{it}=h|H_{it}=\bar{h},z_i), \end{aligned} \tag{7}$$

where a_{hjty} corresponds to the (weighted) frequency of subjects responding by y to the j-th response variable and belonging to latent state h at occasion t, b_{hit} is an indicator variable equal to 1 if subject i belongs to latent class h at occasion t, with $p(H_{i1}=h|z_i)$ being the initial probabilities computed according to equation (2), and $b_{\bar{h}hit}=b_{\bar{h}i,t-1}b_{hit}$ is an indicator variable equal to 1 if the same subject moves from state $\bar{h}$ to state h at occasion t, with $p(H_{it}=h|H_{it}=\bar{h},z_i)$ being the transition probabilities computed according to equation (3).

Since the latent configuration is not known for each individual, the EM algorithm maximizes the log-likelihood above by alternating the following two steps until converge:

- **E-step**: compute the expected value of the frequencies and indicator variables in (7), given the observed data and the current value of the parameters, so as to obtain the expected value of $\ell^*(\boldsymbol{\theta})$;
- **M-step**: update $\boldsymbol{\theta}$ by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ obtained above.

To select the number of latent states k, when this number is not *a priori* known, we rely on the Bayesian Information Criterion (BIC; Schwarz, 1978), which relies on an asymptotic approximation of a suitable transformation of the Bayesian posterior probability of a candidate model. In practice, k is selected on the basis of the observed data through the index

$$BIC = -2\ell(\hat{\boldsymbol{\theta}}) + \log(n)\#\text{par}, \tag{8}$$

where $\ell(\hat{\boldsymbol{\theta}})$ denotes the maximum of the weighted log-likelihood and #par denotes the number of free parameters. Note that, due to the definition of the weights given in (5), the sample size is n also after the application of these weights, as $\sum_{i=1}^{n}\hat{w}_i = n$, and then the penalization term in (8) is correctly specified.

Once parameter estimates have been computed on the basis of the above two-step procedure, we obtain standard errors by a non-parametric bootstrap method (Davison and Hinkley, 1997). This method is

based on re-sampling subjects (with all their observed pre-treatment covariates, treatment, and outcomes) a suitable number of times with replacement from the observed sample and computing the estimate of the parameters for every bootstrap sample. In this way, we suitably take into account the two-step nature of the proposed estimation method.

5 Properties of the proposed approach

In this section we describe the main asymptotic and finite-sample properties about the proposed approach. Asymptotic properties, which are dealt with in the following section, are based on standard results about maximum likelihood estimators, whereas finite-sample properties are studied by simulation.

Before introducing the technical details, it is important recalling that the causal LM model described in Section 3 may be seen as a marginal model in the sense of Robins et al. (2000). In fact, we do not explicitly formulate assumptions on the conditional distribution of the latent variables $H_{it}^{(z)}$ given the pre-treatment covariates $\boldsymbol{X}_i$, but we model the marginal distribution of these variables. On the other hand, in order to study the properties of the proposed approach, and in particular implementing the simulation study, we have to consider a *data generating model* in which the relation between covariates and latent variables is explicit.

5.1 Asymptotic properties

The data generating model is based on the following scheme for $i = 1, \ldots, n$:

1. the vector of covariates $\boldsymbol{x}_i$ is drawn from an unknown distribution $f(\boldsymbol{x})$;
2. given $\boldsymbol{x}_i$, the potential latent outcomes $H_{it}^{(z)}$ are drawn, for $t = 1, \ldots, T$ and $z = 1, \ldots, l$, from a Markov chain based on initial and transition probabilities which arbitrarily depend on the covariates;
3. given $\boldsymbol{x}_i$, the treatment indicator z_i is generated from a multinomial logit model based on the probabilities $p_z(\boldsymbol{x}_i)$, $z = 1, \ldots, l$, and formulated as in equation (4);
4. the latent variables H_{it} are generated as $H_{it} = H_{it}^{(z_i)}$ for $t = 1, \ldots, T$;
5. given the generated value of H_{it}, the outcomes Y_{ijt} are generated, for $j = 1, \ldots, r$ and $t = 1, \ldots, T$, according to the LM model as specified in Section 3.

As already mentioned, the causal LM model that we estimate is a "marginalized" version of the data generating model based on the previous assumptions. Therefore, in order to assess the properties of the

proposed estimator, which is illustrated in Section 4, a crucial issue is determining what is the true value of the parameters of this causal model under the data generating model; the corresponding true parameter vector is denoted by $\boldsymbol{\theta}_0$. A natural way to define $\boldsymbol{\theta}_0$ is as the point of convergence of the standard maximum likelihood estimator of $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ under a "randomized" version of the assumed data generating model; this estimator is obtained from the maximization of the target function

$$\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}). \tag{9}$$

The *randomized sampling scheme* is based on the same steps 1-5 indicated above, but at the third step the treatment indicator variable is drawn from a uniform distribution, which is independent of $\boldsymbol{X}_i$, with probabilities $p_{z0} = 1/l$, $z = 1, \ldots, l$. Note that the proposed estimator $\hat{\boldsymbol{\theta}}$ and the estimator $\tilde{\boldsymbol{\theta}}$ are indeed based on two different sampling schemes, with the first possibly affected by selection bias.

In the following, we establish the consistency of the proposed estimator which is based on the fact that this estimator and the randomized one are based on the maximization of two functions that, divided by n, converge in probability to the same function as the sample size increases.

Proposition 1. *As $n \to \infty$, the estimator $\hat{\boldsymbol{\theta}}$ under the data generating model described above and the estimator $\tilde{\boldsymbol{\theta}}$ under the randomized sampling scheme converge in probability to the same point $\boldsymbol{\theta}_0$ of the parameter space.*

Proof. See Appendix.

A final point we here discuss regards the properties of BIC to select the appropriate number of latent states k. In dealing with standard finite mixture models, the consistency of this criterion has been established in Keribin (2000), where consistency means that the true number of mixture components is selected with probability approaching to 1 as $n \to \infty$. Moreover, McLachlan and Peel (2000) suggests this criterion as suitable in typical applications of mixture models and, similarly, Bacci et al. (2013) showed, on the basis of a deep simulation study, that it tends to outperform alternative selection criteria for the traditional LM model. Moreover, advanced versions of BIC have been recently introduced and studied from the point of view of consistency when the assumed statistical model is estimated by composite, instead of full, composite likelihood; see Gao and Song (2010).

In the present causal framework, the above results are not directly applicable essentially for two reasons. First, the estimated model is not directly the data generating model, but a "marginalized" version of this one. Second, the adopted estimator is not a standard likelihood estimator neither a composite likelihood estimator. However, we expect BIC to perform properly in applications for the causal LM model here

illustrated. This conclusion is supported by the simulation results that are illustrated in the following section.

Regarding the choice of the number of latent states, our point of view is that it is not essential to spot the "correct" number of latent states only on the basis of the observed data in a causal model as the present one. In fact, reasons of interpretability and stability of the results may lead to choosing a different number of latent states with respect to that indicated by a selection criterion. Moreover, adopting solely a selection criterion based on the observed data may lead to choosing an increasing number of states as the number of time occasions T increases. In fact, as T increases, the differences between individuals in terms of observable path tend to increase and it is easier to spot these differences as the amount of information increases. However, the fact that in the same applicative context and for the same group of individuals two different values of k may be suggested, depending on the length of the period of observation, may be undesirable from the interpretative point of view.

5.2 Simulation study

In order to study the finite-sample properties of the estimator illustrated in the previous section, we performed a simulation study in which the estimates obtained from this method are compared with those obtained from the naive estimation method of the LM model without using PS weighting and those obtained in the case of randomized treatment. Details about the simulation design and its results are illustrated in the following.

5.2.1 Simulation design

We assume the existence of two covariates affecting both the treatment and the value of the potential outcomes $H_{it}^{(z)}$ for all possible treatments z; the first of these covariates (x_{i1}) is continuous and is generated from a standard normal distribution, whereas the second (x_{i2}) is binary with two possible values, -0.5 and 0.5, having the same probability. The observed values of these two covariates are collected in the vector $\boldsymbol{x}_i$. The simulation is based on the following generating model for the POs at the initial time occasion

$$\log \frac{p(H_{i1}^{(z)} = h|\boldsymbol{x}_i)}{p(H_{i1}^{(z)} = 1|\boldsymbol{x}_i)} = \alpha_h^* + \beta_{hz}^* + (x_{i1} + x_{i2})\tau_h, \quad h = 2, \ldots, k,\ z = 1, \ldots, l,$$

whereas for the next occasions we have

$$\log \frac{p(H_{it}^{(z)} = h|H_{i,t-1}^{(z)} = \bar{h}, \boldsymbol{x}_i)}{p(H_{it}^{(z)} = 1|H_{i,t-1}^{(z)} = \bar{h}, \boldsymbol{x}_i)} = \gamma_{\bar{h}h}^* + \delta_{hz}^* + (x_{i1} + x_{i2})\psi_h,$$

with $\bar{h} = 1, \ldots, k$, $h = 2, \ldots, k$, and $z = 1, \ldots, l$. Note that in the above expressions we add an asterisk to the parameters α_h, β_{hz}, $\gamma_{\bar{h}h}$, and δ_{hz}, as these parameters are different from those of the causal LM model. In fact, as already explained in Section 5.1, this model does not coincide with the data generating model.

In the base-line simulation design, we assume $k = 2$ latent states and $l = 2$ treatments, and we fix $\alpha_2^* = -1$, $\beta_{22}^* = 2$, $\tau_2 = 1$, $\gamma_{12}^* = -1$, and $\gamma_{22}^* = 1$, with $\delta_{22}^* = \beta_{22}^*/2$ and $\psi_2 = \tau_2/2$, so that the effect of treatment and covariates is smaller on the initial occasion with respect to the following ones. The assignment mechanism of the treatment is based on the assumption

$$\log \frac{p(Z_i = 2|\boldsymbol{x}_i)}{p(Z_i = 1|\boldsymbol{x}_i)} = x_{i1} + x_{i2}.$$

Given the same assignment mechanism, with $k = 3$ latent states, and again $l = 2$ possible treatments, we fix $\alpha_2^* = -0.5$, $\alpha_3^* = -1.5$, $\beta_{22}^* = 1.5$, $\beta_{32}^* = 3$, $\gamma_{12}^* = -1$, $\gamma_{13}^* = -2$, $\gamma_{22}^* = 1$, $\gamma_{23}^* = 0$, $\gamma_{32}^* = 1$, $\gamma_{33}^* = 2$, $\tau_2 = 0.5$, and $\tau_3 = 1$, with $\delta_{h2}^* = \beta_{h2}^*/2$ and $\psi_h = \tau_h/2$ for $h = 2, 3$.

Under the scenario with $l = 3$ treatments, we assume that the treatment is assigned according to the multinomial logit model based on the following assumption:

$$\log \frac{p(Z_i = z|\boldsymbol{x}_i)}{p(Z_i = 1|\boldsymbol{x}_i)} = (z-1)(x_{i1} + x_{i2}), \quad z = 2, 3.$$

Moreover, with $k = 2$ we use the same values as above for all the parameters of the model for the potential outcomes, with $\beta_{22}^* = 1$ and $\beta_{23}^* = 2$. Similarly, with $k = 3$ we assume $\beta_{22}^* = 0.75$, $\beta_{23}^* = 1.5$, $\beta_{32}^* = 1.5$, and $\beta_{33}^* = 3$. In both cases we have $\delta_{hz}^* = \beta_{hz}^*$ for all h and z.

Under each model defined as above, we drew 1,000 samples of size $n = 1000, 2000$ for a number of time occasions $T = 4, 8$, so that there are 16 scenarios overall. For every sample, the causal effects were estimated by the proposed method and its unweighted version. Moreover, for comparison, the corresponding estimates based on a randomized design were obtained. The latter amounts to draw the assignment variables Z_i from a generalized Bernoulli distribution with l categories having the same probability, and then independently from the pre-treatment covariates, and applying the unweighted estimator on the resulting data.

5.2.2 Simulation results

The results in terms of parameter estimates are reported in Table 1 for the case of $l = 2$ treatments. We also considered selection of k on the basis of BIC. Regarding this aspect of model selection, the results are reported in Table 2 for $l = 2$. For $l = 3$, the results are in Tables 3 and 4. In Tables 1 and 3, the estimator based on randomized samples is denoted as "randomized", that based on the proposed estimation

approach when samples are affected by confounding is denoted as "proposed", and the estimator which does not correct for confounding is denoted as "naive", being based on the maximization of the standard (unweighted) likelihood function.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

From the results in Tables 1 and 3, which are obtained under different scenarios, we conclude that the bias of the proposed estimator is typically negligible. We stress that this bias is computed as the difference between the mean of the estimates based on the proposed method and that of the estimates based on the randomized experiment, as we do not directly estimate the parameters of the data generating model. Therefore, it would not be correct to directly compare, for instance, the average of the estimates of the parameters β_{hz} with the proposed method with the true value of the parameters β^*_{hz} assumed in the data generating model. We also find that, as expected, the standard deviation of the proposed estimator decreases as n and T increase; in particular, it decreases with a rate close to $\sqrt{n}$.

On the other hand, the bias is typically large for the naive estimator. This bias does not reduce as the sample size and/or the number of time occasions increases and confirms that, differently from the proposed estimator, the naive estimator is inadequate and may lead to completely wrong results. These results would also be very different from those obtainable if the treatment was perfectly randomized.

Finally, it is worth noting that, according to the results in Tables 2 and 4, BIC performs very well in choosing the optimal number of latent states. Therefore, we conclude that it is an adequate selection criterion also when the weighted log-likelihood estimator is applied and this justifies the rule in equation (5) to compute the PS weights.

6 Application

In this section, the proposed methodology is illustrated through an application focused on the estimation of ATE of the degree type on the Human Capital (HC) development. In particular, we compare individuals who graduated in different subjects taking into account the university labour market transition in the first period after graduation, so as to consider their work path; see also Bartolucci and Pennoni (2011).

In the following, we first describe the context of the study and the available data. Then, we show the main results of the application and we provide some remarks about the estimated effectiveness of the different types of degree.

6.1 Data description

We represent HC by a sequence of individual latent variables having a longitudinal dimension, so as to study its evolution across time. This follows Heckman (2000), who conceived HC as a latent variable which is affected by investment in education, individual ability, type of higher education. In turns, HC affects income, other outcomes (manifest variables) describing the stability of the job position, and skills actually employed in comparison with the acquired ones.

In order to properly assess the effect of university studies on the transition to the labor market, we consider the following aspects: (*i*) different outcomes may describe how much knowledge improves the personal opportunities on the job market, taking into account not only earnings but also indicators of job stability and improvement in skills; in this regard, Harpan and Draghici (2014) suggested that HC must be measured also including non-monetary aspects as well as considering its development; (*ii*) the phenomenon has a longitudinal perspective and the evolution of the outcomes observed at repeated occasions is of main importance; (*iii*) the treatment of subjects is not controlled and there are multiple potential treatments corresponding to different types of degree program. Also note that, as in other model frameworks for causal inference, the pre-treatment covariates are of main importance as the treatment groups differ, already prior to treatment, in a way that can affect the outcomes. This clearly happens in our application so that the groups of individuals with different degrees need to be balanced on the basis of these covariates.

The available data derive from certain administrative archives and concern graduates from the Lombardy region. The Supplementary Material file of the present article provides more details on the administrative archives from which the data are obtained. The dataset refers to 1,144 individuals who graduated in 2007 from four universities and are resident in the area surrounding Milan. Note that, in this area, these universities are comparable in terms of prestigious and quality and then we rule out a differential effect between them in terms of work path. On the other hand, we study the effect of different types of academic graduation; in other words, the type of graduation is the treatment of interest. For this aim, we deal with a cohort of graduates who have been observed along four quarters after graduation, covering in this way one year.

The percentage of the graduates for each type of degree, as well as the descriptive statistics for the available pre-treatment covariates conditional on each treatment, are provided in the Supplementary Material

file. In this regard, note that Z_i is discrete variable with $l = 5$ treatments equal to: 1 for technical degrees, 2 for architecture, 3 for economic degrees, 4 for humanities degrees, and 5 for scientific degrees. The groups are not balanced in terms of pre-treatment covariates; in particular, there are relevant differences between graduates with a technical degree and those with other degrees.

The following $r = 3$ response variables concerning graduates' employment status are available for each of the four quarters of observation:

i) *contract type* with categories: none, temporary, and permanent;

ii) *skill* with categories: none, low/medium, and high;

iii) *gross income* in Euros (€) with categories: none, ⩽3750 €, and >3750 €.

Then, we have $r = 3$ and $c_j = 3$ for $j = 1, \ldots, r$, as all variables have three response categories. Note that the category none means that the contract type is not temporary neither permanent but some less qualified type of contract. The category none for skill corresponds to those jobs not requiring any qualified skill, and for income this category refers to incomes gained from other sources.

Regarding the response variables, note that we essentially distinguish between temporary and permanent contracts. Moreover, the gross income is reported quarterly and we choose to consider the threshold of €15,000 yearly, which corresponds to €3,750 quarterly. The third response variable is based both on skill level and skill specialization. According to the definition of HC given above, the observed response variables we dispose represent a manifestation of this latent construct. The local independence assumption is reasonable and, from our point of view, no other dependences in the data need to be specified.

6.2 Results

We first consider the selection of the pre-treatment covariates to be used to compute the weights (first step). Then, we show the estimated parameters related to ATE of the degree on the work path (second step).

6.2.1 Selection of the pre-treatment covariates

As illustrated in Section 4, the individual weights are computed on the basis of the estimated multinomial logit model for the treatment, which is based on equation (4).

In general, it is important to select the appropriate pre-treatment covariates that enter in this model; see also McCaffrey et al. (2013). In this respect, we start from an analysis of the dependence of each of these covariates with the type of university degree. For quantitative covariates, such as family's income,

the analysis is based on an ANOVA model and only covariates for which a clear dependence with the type of degree is ascertained are included in the multinomial logit model. covariate is considered to have a strong dependence with the choice of the degree type when the hypothesis is rejected that the means of this covariate for the different degrees are equal. In the case of qualitative covariates, such as gender, we instead use a chi-square test of independence for contingency tables. Therefore, we choose the following among the pre-treatment covariates: *gender*, *district of birth*, *final grade at high school diploma*, *type of high school*. The parameter estimates of this multinomial model are reported in Table 5.

[Table 5 about here.]

In Table 6 we report some descriptive statistics for the distribution of each of these covariates given the degree type and accounting for the PS weights based on the fitted multinomial model.

[Table 6 about here.]

From Table 6 we observe that the balance between the groups corresponding to the different university degrees is considerably higher with respect to the initial distributions (compare with Table 2 in the Supplementary Material file). For instance, the percentage of females among graduates in a technical subject is around 20% and it is around 80% among graduates in a humanities subject. After re-weighting, these percentages become equal to 45% and 49%, respectively, and they are close to those computed for the other university degrees. A similar convergence between the distributions corresponding to the other degree types is observed for the remaining three covariates.

6.2.2 Results of the model fitting

After the individual PS weights have been obtained, we estimated the proposed LM model for an increasing number of latent states not larger than 4, so as to avoid models of difficult interpretation, and under this constraint we selected this number on the basis of BIC, as illustrated in Section 4. It turns out that the number of latent classes is $k = 4$ corresponding to the minimum of the BIC index among the values of k we considered. The corresponding maximum weighted log-likelihood is $\ell(\hat{\boldsymbol{\theta}}) = -5180.00$, with 63 free parameters, so that $BIC = 10803.66$.

The estimates of the conditional response probabilities ($\phi_{jy|h}$) obtained under the selected model with four latent states are reported in Table 7 and are also illustrated in Figure 1.

[Table 7 about here.]

[Figure 1 about here.]

According to the estimates of the parameters $\phi_{jy|h}$ we can rather easily interpret the latent states according the corresponding distribution of the response variables. In particular, the first class corresponds to the cluster of unemployed individuals, and then the probability of the first category is equal to 1 for each response variable. On the other hand, the fourth latent class corresponds the subpopulation of individuals in very good work conditions, as they have a permanent contract and tend to have a job with high skill level and high income. The interpretation of the other two classes is less straightforward, even if it is clear that they are intermediate between the first class and the fourth. In particular, the second class is that of individuals that typically have a temporary contract for a low/medium skill job. Individuals in the third class differ from those in the second for having a higher probability of making use of their high skills. Overall, the interpretation of these classes in terms of HC may be summarized as follows:

- *1st class*: lowest HC level;
- *2nd class*: intermediate HC level with high probability of a temporary job, requiring a low/medium skill level, and intermediate income level;
- *3rd class*: similar to the 2nd class with the exception of a higher skill level and a slightly higher income;
- *4th class*: high HC level with permanent contract, intermediate-high skills, and high income.

It is also worth noting that, although these latent classes can be easily interpreted, they are non-monotonically ordered, implying that the constraint of monotonicity only holds for certain pairs of classes, such as the pair (1,4); however, this constraint does not hold for the pair (3,4). Then, avoiding to incorporate in the model the constraint of monotonicity by parametrizing the conditional response probabilities is a proper choice as this constraint would reduce the model fit considerably.

The estimates of the regression coefficients α_h and β_{hz} affecting the initial probabilities of the hidden Markov chain, see equation (2), are reported in Table 8, with the indication of the significance level for the test that each of these parameters is equal to 0. For the corresponding estimated initial probabilities see Table 9. We recall that every parameter β_{hz} may be interpreted as an ATE of degree (treatment) z, with respect to a technical degree ($z = 1$), in terms of initial probabilities expressed on the logit scale. In Table 8 we also report the estimates of the pairwise differences $\beta_{hz_1} - \beta_{hz_2}$, $z_2 \neq z_1$, which can be interpreted in terms of ATE of treatment z_2 with respect to treatment z_1. The estimates of the parameters $\gamma_{\bar{h}h}$ and δ_{hz} referred to the causal effect on the transition probabilities, see equation (3), are reported in Table 10, together with estimates of the differences $\gamma_{hz_1} - \gamma_{hz_2}$, $z_1 = 2, \ldots, l-1$, $z_2 \neq z_1$, and the indication of the

significance level for each of these parameters, which may be again interpreted in terms of ATE on the transition probabilities.

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

On the basis of the estimates of the logit regression parameters for the initial probabilities, reported in in Table 8, we conclude that, at the beginning of the period of observation, there is a statistical significant difference in terms of effect on HC of technical degrees with respect to architecture and humanities degrees and in favor of the first ones. There is also a significant differential effect of economic degrees with respect to architecture and humanities degrees. This conclusion is confirmed by the estimated initial probabilities for each type of treatment given in Table 9, which show that technical and economic degrees have the lowest probability that a graduate is in the first latent class (0.45 for both degrees) and the highest probability that he/she in the last latent class (0.17 for technical and 0.15 for economic degree). Then, scientific degrees correspond to intermediate estimates of both probabilities (0.65 for the first class and 0.10 for the last) whereas, with reference to architecture and humanities degrees, we have the highest probabilities for the first class (0.75 and 0.67, respectively) and the lowest probabilities for the last class (0.07 and 0.05, respectively). It is interesting to note that, looking at the initial probabilities of the first or the last class, we obtain substantially the same ranking of the degree types in terms of effect of initial HC level, an aspect that gives consistency to our findings. In summary, we can rank the degrees as follows in terms of their effectiveness at the beginning of the period after graduation: technical, economic, scientific, humanities, architecture. We rank architecture as the last degree type on the basis of the initial probability of the first latent class although humanities has a lower probability for the last latent class, but the difference is small.

The picture is somehow different in terms of causal effects of the degree type on the evolution of the HC level, which is represented through transition probabilities. In fact, on the basis of the estimates of the logit regression parameters affecting the transition probabilities (Tables 10) we conclude that there are significant differences between technical degrees and all the other types of degree during the period of observation and between economic and architecture degrees in favor of the first.

The above conclusions in terms of ranking of the degree types are confirmed by the estimated transition probability matrices reported in Table 11 for each type of degree and which are computed on the basis of the estimates reported in Table 10.

[Table 11 about here.]

All matrices are characterized by a rather high persistence, with elements in the main diagonal always greater than 0.65, many of which are also greater than 0.8. However, some differences emerge between these transition matrices. In order to ease the comparison, it is again convenient to focus on the first and the last latent classes, corresponding to the subpopulations of individuals with the lowest and the highest HC level, respectively. In particular, we observe that for technical degrees there is the lowest probability of remaining in the first latent class (0.72) and then the highest probability of moving away from this class, meaning that this type of degree induces the best improvement in terms of HC. For this type of degree we also have the highest probability of remaining in the last class (1.00). The second lowest probability of remaining in the first class is for economic degrees (0.84), corresponding to the second highest probability of remaining the last class (0.99). The third lowest probability of remaining in the first class is for humanities degrees (0.85), corresponding to the second lowest probability of remaining in the last class (0.97). Finally, for scientific degrees we have the second lowest probability of persistence in the first latent class (0.86) and the third lowest probability of remaining in the last class (0.97), whereas for architecture we have the highest probability for the first class (0.89) and the lowest for the last class (0.95).

Overall, we conclude that there is a rather clear ranking between university degrees in terms of impact on the HC level which is as follows:

- *Technical degrees*: highest effect at the beginning and in terms of evolution of HC level;
- *Economic degrees*: impact close to technical degrees and in terms of evolution of HC level;
- *Scientific degrees*: significantly worse impact with respect to technical and economic degrees;
- *Humanities and Architecture degrees*: worse effect at the beginning and in terms of evolution of HC level.

In order to assess how these conclusions are sensitive to the initially selected set of covariates, we performed the same analysis using all the available pre-treatment covariates to compute the individual weights. This amounts to fit the multinomial logit model, based on equation (4), with all these covariates and obtain in this way the PS weights. The results of this second fitting are very close to those obtained above, leading to the same conclusions about the effect of the treatment of interest. For details about this comparison we refer the reader to the second part of the Supplementary Material file.

7 Conclusions

We propose a novel approach for estimating Average Causal Effects (ATEs) when dealing with longitudinal data in observational studies and in the presence of multiple treatments. It is based on integrating the Latent Markov (LM) model with modeling techniques based on the potential outcome framework (Rubin, 1974, 2005). We introduce a causal inference perspective into the LM model and, at the same time, we extend the causal inference approach developed by Lanza et al. (2013) to the longitudinal context. This innovative statistical method has a potential use in a wide range of observational studies which rely on the same assumptions. It allows us to summarize the multivariate responses observed at each occasion by latent classes (or states) having the interpretation of clusters (or subpopulations) of individuals. The flexibility of the adopted parameterization allows us to deal with any kind of response variable, not only categorical. Of main importance is the estimation of the ATEs which are expressed in terms of initial probabilities of the latent classes and transition probabilities between these classes, so as to separate the impact at the beginning of the period of observation from that on the evolution of the characteristic of interest.

The model is fitted by a two-step maximum likelihood estimation procedure based on first estimating a multinomial logit model for the probability of taking each type of treatment given suitably chosen pre-treatment covariates. Then, a weighted log-likelihood of the LM model, with weights computed on the basis of the estimates computed at the first step, is maximized so as to obtain the parameter estimates. This second step is based on the expectation-maximization algorithm (Dempster et al., 1977). Reliable standard errors for the model parameters are obtained by using a non-parametric bootstrap method (Davison and Hinkley, 1997). The number of latent classes is selected by the Bayesian Information Criterion (BIC; Schwarz, 1978).

The selection of the appropriate pre-treatment covariates is made by considering an ANOVA model for the quantitative covariates and a chi-square test of independence for qualitative covariates. These covariates are selected if they show a significant dependence with a specific type of treatment. Moreover, the use of the multinomial logit parameterization for the initial and for the transition probabilities of the latent variables implies that the number of parameters to be estimated may be potentially high. However, we introduce a constrained form that makes the model more parsimonious and, in the causal context, this is a suitable parameterization because it allows us to estimate the effects of interest in case of multiple treatments. A partially ordered hidden Markov model, as that proposed by Ip et al. (2013), may be also considered.

We assess the asymptotic properties of the proposed estimator by taking into account the data generating model. We provide a proposition of its consistency by showing that it converges to the same point of the parameter space at which the standard estimator converges under a perfectly randomized sampling scheme. Then, we assess the finite-sample properties of the proposed causal effect estimator by means of a simulation

study. In this study, we compare the results especially in terms of bias of the proposed weighted estimator with the naive unweighted estimator of the LM model and with the estimator based on a randomized assignment of the treatment. As expected, the proposed estimator has a negligible bias, whereas the naive estimator may have a huge bias. We also evaluate the performance of the BIC and we conclude that this criterion is able to select the correct number of latent states also when used within the weighted likelihood approach. According to the simulation study we conclude that the proposed approach leads us to an adequate estimator of the causal effects of interest.

In the application, we aim at studying the development of the Human Capital (HC), and the related university-to-work transition phenomenon, due to the different types of degree. We refer to a definition of HC as a latent construct related to the skills, competencies, and attributes embodied in individuals that are relevant to the economic activities, with particular reference of the labor market. The response variables are categorical and they have been chosen in order to capture different aspects of the individuals. In fact, we conceive HC as a "potential version" of a latent variable which underlies the three employment quality measures, in accordance with the most recent definitions of HC. Therefore, the local independence assumption is reasonable. The analyzed data are referred to the first year after graduation of the entire population of graduates in 2007 in four universities in Milan and certain important pre-treatment covariates are available. We observe the relevant factors contributing to the treatment assignment and then the assumption of absence of unmeasured confounders is no doubtful. However, the latter may be also assessed by a sensitivity analysis which may be considered in subsequent studies. By applying the proposed approach to the data at hand, we selected a model with four latent states. We show that the results have an easy interpretation and a rather insensitive to the specification of the multinomial logit model used to obtain the initial Propensity Score (PS) weights. We conclude that the different types of academic degree have significantly different effects on the work path. The choice of different treatments (type of degree) has an impact on the return of investment in HC, the increase of which is higher for those having a technical and economic degrees respect to other degrees. The model also allows us to rank university degrees in terms of impact on HC levels. It also worth mentioning that, by applying the same proposal and relaying on more time occasions, we might study the long term effects of the degrees.

Finally, we stress that the proposed model is very flexible as it may be easily extended to the case of mixed response variables. Another possible extension of interest is for multilevel data; this extension could be formulated according to the approach adopted in Bartolucci et al. (2011) where additional discrete latent variables are introduced to account for the dependence between the individuals in the same clusters. In this case a two-step estimator may be still used, with the first step consisting in the computation of the

PS weights and the second consisting in the maximization of the weighted model likelihood. The main complication is in this second step and, in particular, in the estimation algorithm. This difficulty is given by the presence of a complex latent structure based on several latent variables that are dependent each other. Moreover, special care must be paid to the method for obtaining the standard errors for the parameter estimates, so as to account for the within-cluster dependence in addition to the two-step nature of the estimator. As suggested by a Referee, one possibility would be to implement a bootstrap algorithm based on resampling units within each cluster.

Acknowledgments

We thank the Editor and the Reviewers for many stimulating comments. We are also grateful to Prof. M. Mezzanzanica and to Dr. M. Fontana, of the Interuniversity Research Centre on Public Services (CRISP), University of Milano-Bicocca, for providing the analyzed dataset. Finally, we acknowledge the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project "Mixture and latent variable models for causal inference and analysis of socio-economic data"). F. Pennoni also thanks the financial support of the STAR project "Statistical models for human perception and evaluation" funded by the University of Naples Federico II.

Appendix

Proof of Proposition 1. First of all, consider the average log-likelihood defined according to (6), that is,

$$\bar{\ell}(\boldsymbol{\theta}) = \frac{\ell(\boldsymbol{\theta})}{n} = \frac{\sum_{i=1}^{n} 1/\hat{p}(z_i|\boldsymbol{x}_i) \log p(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}|z_i)}{\sum_{i=1}^{n} 1/\hat{p}(z_i|\boldsymbol{x}_i)}.$$

Since the parameters of the logit model are consistently estimated at the first step, then also each $\hat{p}(z_i|\boldsymbol{x}_i)$ is uniformly consistently estimated and this implies that

$$\bar{\ell}(\boldsymbol{\theta}) \xrightarrow{p} \frac{\mathrm{E}[1/p(Z_i|\boldsymbol{X}_i) \log p(\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}|Z_i)]}{\mathrm{E}[1/p(Z_i|\boldsymbol{X}_i)]}, \tag{10}$$

where the expected value $\mathrm{E}(\cdot)$ is computed with respect to the joint distribution of $\boldsymbol{X}_i$, Z_i, and $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}$, under the data generating model; this distribution does not depend on a specific individual i.

The denominator of (10) may be rewritten as

$$\mathrm{E}[1/p(Z_i|\boldsymbol{X}_i)] = \sum_{z=1}^{l} \left\{ \int_{\boldsymbol{x}} [1/p(Z_i = z|\boldsymbol{X}_i = \boldsymbol{x})] p(Z_i = z|\boldsymbol{X}_i = \boldsymbol{x}) f_0(\boldsymbol{x}) d\boldsymbol{x} \right\} = l,$$

where $f_0(\boldsymbol{x})$ refers to the true distribution of the covariates. Using similar arguments, we also have that

$$\begin{aligned}
&\mathrm{E}[1/p(Z_i|\boldsymbol{X}_i) \log p(\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}|Z_i)] \\
&= \sum_{z=1}^{l} \int_{\boldsymbol{y}_1} \cdots \int_{\boldsymbol{y}_T} f_0(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T|z) \log p(\boldsymbol{Y}_{i1} = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_{iT} = \boldsymbol{y}_T|Z_i = z) d\boldsymbol{y}_T \cdots d\boldsymbol{y}_1,
\end{aligned}$$

where

$$f_0(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T|z) = \int_{\boldsymbol{x}} p(\boldsymbol{Y}_{i1} = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_{iT} = \boldsymbol{y}_T|\boldsymbol{X}_i = \boldsymbol{x}, Z_i = z) f_0(\boldsymbol{x}) d\boldsymbol{x}$$

is the conditional distribution of $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}$ given Z_i, provided that Z_i is independent of $\boldsymbol{X}_i$. Consequently, we have

$$\bar{\ell}(\boldsymbol{\theta}) \xrightarrow{p} \mathrm{E}_0[\log p(\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}|Z_i)], \quad \forall \boldsymbol{\theta},$$

where the expected value $\mathrm{E}_0(\cdot)$ is computed under the randomized sampling scheme in which each treatment has the same probability $1/l$, so that

$$\begin{aligned}
&\mathrm{E}_0[\log p(\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}|Z_i)] \\
&= \frac{1}{l} \sum_{z=1}^{l} \int_{\boldsymbol{y}_1} \cdots \int_{\boldsymbol{y}_T} f_0(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T|z) \log p(\boldsymbol{Y}_{i1} = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_{iT} = \boldsymbol{y}_T|Z_i = z) d\boldsymbol{y}_T \cdots d\boldsymbol{y}_1. \qquad (11)
\end{aligned}$$

Now consider the average version of the target function defined in equation (9), that is,

$$\tilde{\ell}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log p(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}|z_i).$$

By standard arguments we have that

$$\tilde{\ell}(\boldsymbol{\theta}) \xrightarrow{p} \mathrm{E}_0[\log p(\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}|Z_i)], \quad \forall \boldsymbol{\theta}.$$

Therefore, the two target functions on which the two estimators are based convergence to the same function defined in equation (11). Since usual regularity conditions about the two involved likelihoods are satisfied, it follows that both $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ converges in probability to the same point $\boldsymbol{\theta}_0$ of the parameter space. □

References

Aalen, O. O., Roysland, K., Gran, J. M., and Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A*, 175:831–861.

Angrist, J. D. (1991). Grouped-data estimation and testing in simple labor-supply models. *Journal of Econometrics*, 47:243–266.

Arjas, E. (2013). Time to consider time, and time to predict? *Statistics in Biosciences*, 6:1–15.

Bacci, S., Pandolfi, S., and Pennoni, F. (2013). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, 8:125–145.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, series B*, 68:155–178.

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, 104:816–831.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton, FL.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates (with discussion). *Test*, 23:433–486.

Bartolucci, F. and Pennoni, F. (2011). Impact evaluation of job training programs by a latent variable model. In *New Perspectives in Statistical Modeling and Data Analysis*, pages 65–73. Springer-Verlag, Berlin.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioural Statistics*, 36:491–522.

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

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, MA.

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

Gao, X. and Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105:1531–1540.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Guo, S. and Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications*. Sage, Thousand Oaks, CA.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer Nijhoff, Boston.

Harpan, I. and Draghici, A. (2014). Debate on the multilevel model of the human capital measurement. *Procedia - Social and Behavioral Sciences*, 124:170 – 177.

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics*, 54:3–56.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96:440–448.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87:706–710.

Ip, E., Zhang, Q., Rejeski, J., Harris, T., and Kritchevsky, S. (2013). Partially ordered mixed hidden Markov model for the disablement process of older adults. *Journal of the American Statistical Association*, 108:370–384.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya: The Indian Journal of Statistics, Series A*, 62:49–66.

Lanza, S. T., Coffman, D. L., and Xu, S. (2013). Causal inference in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 20:361–383.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32:3388–3414.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42:109–142.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Pennoni, F. (2014). *Issues on the Estimation of Latent Variables and Latent Class Models*. Scholars' Press, Saarbucken.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331.

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

Vermunt, J. K., Langeheine, R., and Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24:179–207.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:1–13.

Wiggins, L. (1955). Mathematical models for the analysis of multi-wave panels. In *Ph.D. Dissertation*, Ann Arbor. Columbia University.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Springer-Verlag, New York.

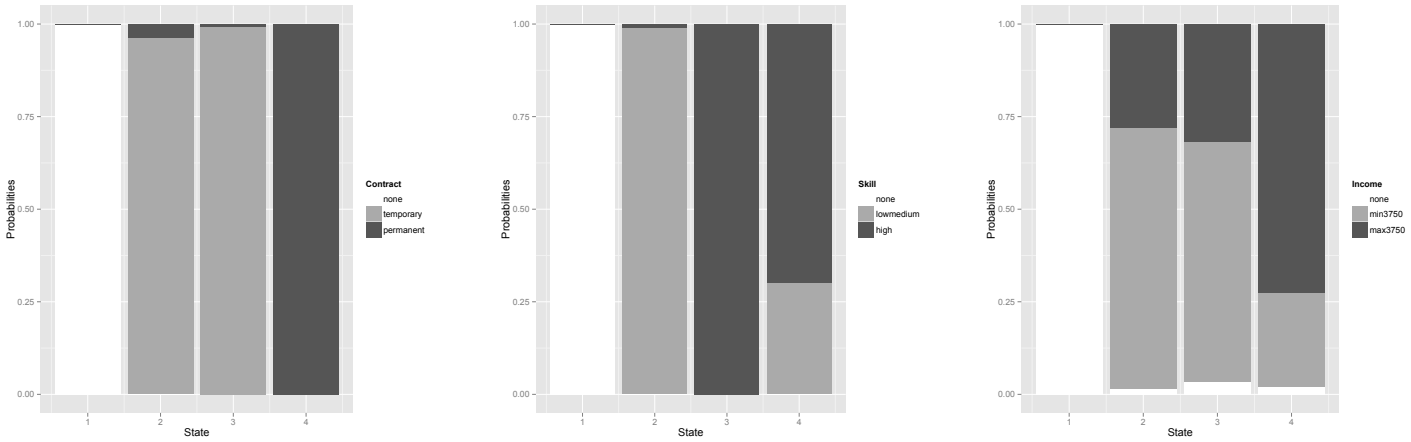

Figure 1: *Estimated conditional probabilities of labor condition ($\phi_{jy|h}$), under the proposed causal LM model with $k = 4$ latent states.*

				$k=2$		$k=3$			
				β_{22}	δ_{22}	β_{22}	β_{23}	δ_{22}	δ_{23}
$n=1000$	$T=4$	randomized	mean	1.622	0.798	1.308	2.591	0.634	1.278
			sd	0.254	0.170	0.261	0.252	0.172	0.168
		proposed	mean	1.641	0.790	1.297	2.623	0.644	1.290
			bias	0.019	-0.008	-0.011	0.032	0.011	0.013
			sd	0.324	0.228	0.349	0.345	0.261	0.232
		naive	mean	2.538	1.250	1.786	3.586	0.882	1.771
			bias	0.916	0.453	0.478	0.995	0.248	0.494
			sd	0.306	0.194	0.295	0.298	0.209	0.187
	$T=8$	randomized	mean	1.613	0.819	1.280	2.590	0.652	1.307
			sd	0.239	0.095	0.261	0.252	0.110	0.102
		proposed	mean	1.616	0.824	1.310	2.626	0.652	1.315
			bias	0.003	0.005	0.030	0.037	0.001	0.008
			sd	0.306	0.129	0.346	0.330	0.158	0.141
		naive	mean	2.510	1.283	1.795	3.581	0.888	1.794
			bias	0.897	0.464	0.515	0.991	0.237	0.487
			sd	0.287	0.113	0.301	0.302	0.127	0.114
$n=2000$	$T=4$	randomized	mean	1.606	0.796	1.288	2.586	0.631	1.274
			sd	0.167	0.116	0.196	0.175	0.123	0.114
		proposed	mean	1.606	0.800	1.281	2.579	0.645	1.290
			bias	0.000	0.004	-0.006	-0.007	0.014	0.016
			sd	0.219	0.151	0.240	0.230	0.178	0.160
		naive	mean	2.522	1.253	1.769	3.546	0.879	1.768
			bias	0.917	0.457	0.481	0.959	0.248	0.493
			sd	0.211	0.135	0.196	0.205	0.146	0.134
	$T=8$	randomized	mean	1.596	0.817	1.284	2.573	0.646	1.302
			sd	0.168	0.065	0.177	0.163	0.078	0.071
		proposed	mean	1.598	0.817	1.291	2.578	0.649	1.305
			bias	0.002	0.000	0.007	0.006	0.003	0.003
			sd	0.209	0.087	0.239	0.232	0.113	0.098
		naive	mean	2.507	1.275	1.775	3.544	0.889	1.788
			bias	0.910	0.457	0.491	0.971	0.243	0.485
			sd	0.199	0.077	0.194	0.199	0.092	0.082

Table 1: *Results in terms of parameter estimates for the case of $l=2$ treatments.*

.

			selected k				
			1	2	3	4	$\geqslant 5$
$k=2$	$n=1000$	$T=4$	0	972	21	7	0
		$T=8$	0	966	23	7	4
	$n=2000$	$T=4$	0	990	9	1	0
		$T=8$	0	982	15	1	2
$k=3$	$n=1000$	$T=4$	0	0	983	15	2
		$T=8$	0	0	977	17	6
	$n=2000$	$T=4$	0	0	990	9	1
		$T=8$	0	0	983	13	4

Table 2: *Results in terms of selection of k for the case of $l=2$ treatments.*

.

				$k=2$				$k=3$							
n	T			β_{22}	β_{23}	δ_{22}	δ_{23}	β_{22}	β_{23}	β_{32}	β_{33}	δ_{22}	δ_{23}	δ_{32}	δ_{33}
1000	4	rand.	mean	0.821	1.631	0.390	0.799	0.644	1.291	1.315	2.603	0.314	0.641	0.642	1.291
			sd	0.288	0.301	0.190	0.208	0.260	0.328	0.265	0.298	0.170	0.218	0.172	0.207
		prop.	mean	0.822	1.630	0.399	0.812	0.643	1.301	1.336	2.650	0.320	0.649	0.643	1.293
			bias	0.001	-0.001	0.009	0.013	0.000	0.010	0.021	0.047	0.006	0.008	0.000	0.002
			sd	0.332	0.392	0.201	0.260	0.304	0.413	0.333	0.399	0.193	0.300	0.194	0.272
		naive	mean	1.293	2.574	0.628	1.275	0.897	1.802	1.825	3.627	0.441	0.882	0.891	1.785
			bias	0.472	0.943	0.237	0.476	0.253	0.510	0.510	1.024	0.127	0.241	0.249	0.494
			sd	0.317	0.374	0.182	0.225	0.278	0.360	0.304	0.349	0.182	0.253	0.183	0.235
	8	rand.	mean	0.820	1.629	0.405	0.821	0.641	1.301	1.300	2.591	0.327	0.653	0.655	1.306
			sd	0.268	0.293	0.102	0.116	0.270	0.337	0.257	0.291	0.110	0.134	0.108	0.128
		prop.	mean	0.838	1.645	0.404	0.825	0.643	1.296	1.307	2.623	0.319	0.657	0.655	1.321
			bias	0.018	0.016	-0.001	0.004	0.002	-0.005	0.007	0.032	-0.008	0.004	0.000	0.015
			sd	0.331	0.370	0.110	0.148	0.284	0.419	0.324	0.397	0.115	0.195	0.119	0.175
		naive	mean	1.303	2.575	0.637	1.299	0.894	1.792	1.801	3.611	0.444	0.903	0.902	1.818
			bias	0.482	0.947	0.232	0.477	0.253	0.492	0.500	1.020	0.117	0.250	0.247	0.512
			sd	0.317	0.360	0.108	0.129	0.267	0.355	0.304	0.363	0.110	0.158	0.115	0.145
2000	4	rand.	mean	0.789	1.598	0.402	0.793	0.640	1.285	1.302	2.600	0.321	0.641	0.640	1.282
			sd	0.191	0.205	0.123	0.140	0.190	0.226	0.184	0.206	0.117	0.151	0.119	0.145
		prop.	mean	0.807	1.608	0.398	0.799	0.654	1.300	1.303	2.596	0.323	0.640	0.641	1.290
			bias	0.017	0.010	-0.004	0.006	0.015	0.015	0.001	-0.004	0.002	0.000	0.001	0.007
			sd	0.221	0.252	0.137	0.173	0.209	0.277	0.223	0.271	0.135	0.205	0.136	0.188
		naive	mean	1.274	2.548	0.632	1.269	0.907	1.801	1.798	3.590	0.446	0.883	0.888	1.781
			bias	0.485	0.950	0.231	0.476	0.267	0.516	0.496	0.990	0.125	0.243	0.248	0.499
			sd	0.217	0.242	0.129	0.150	0.196	0.240	0.204	0.248	0.127	0.171	0.127	0.157
	8	rand.	mean	0.799	1.605	0.404	0.815	0.642	1.281	1.279	2.567	0.317	0.649	0.648	1.308
			sd	0.191	0.201	0.074	0.081	0.193	0.212	0.175	0.201	0.075	0.097	0.078	0.092
		prop.	mean	0.794	1.592	0.407	0.816	0.644	1.278	1.295	2.583	0.321	0.650	0.648	1.308
			bias	-0.005	-0.013	0.003	0.001	0.002	-0.003	0.016	0.016	0.004	0.001	0.000	-0.001
			sd	0.210	0.243	0.075	0.106	0.208	0.292	0.221	0.264	0.080	0.134	0.076	0.118
		naive	mean	1.262	2.531	0.639	1.290	0.891	1.772	1.786	3.571	0.444	0.896	0.895	1.806
			bias	0.463	0.926	0.235	0.475	0.250	0.492	0.508	1.005	0.127	0.247	0.247	0.498
			sd	0.217	0.232	0.072	0.091	0.192	0.251	0.204	0.237	0.077	0.110	0.076	0.099

Table 3: *Results in terms of parameter estimates for the case of $l = 3$ treatments.*

.

			selected k				
			1	2	3	4	$\geqslant 5$
$k=2$	$n=1000$	$T=4$	0	995	4	1	0
		$T=8$	0	988	9	2	1
	$n=2000$	$T=4$	0	997	2	1	0
		$T=8$	0	993	4	1	2
$k=3$	$n=1000$	$T=4$	0	0	994	5	1
		$T=8$	0	0	988	12	0
	$n=2000$	$T=4$	0	0	998	1	1
		$T=8$	0	0	994	4	2

Table 4: *Results in terms of selection of k for the case of $l = 3$ treatments.*

.

Covariate		Degree (vs. technical)			
		arch.	econ.	human.	scien
intercept		5.677**	6.061**	4.932**	2.074*
gender:	female	1.878**	1.628**	3.156**	1.323**
district of birth:	Lombardy	0.139	-0.338	-0.762†	-1.182
	Italy	-0.507†	-0.311	-0.297	-0.867†
	others	0.853	-1.299†	-0.130	0.007
final score high school diploma:		-0.083**	-0.086**	-0.076**	-0.044**
type of high school:	others	0.799**	1.497**	0.694**	-0.091

Table 5: *Parameter estimates of the multinomial logit model to compute the individual weights (the reference category for each categorical covariate is the one not listed; †significant at 10%, *significant at 5%, **significant at 1%).*

Covariate		University degree				
		techn.	arch.	econ.	human.	scien.
gender:	male	0.552	0.525	0.504	0.514	0.536
	female	0.448	0.475	0.496	0.486	0.464
district of birth:	Milan	0.784	0.726	0.762	0.750	0.791
	Lombardy	0.055	0.060	0.060	0.079	0.059
	Italy	0.146	0.185	0.145	0.147	0.127
	others	0.015	0.029	0.032	0.024	0.023
final score high school diploma:		81.77	81.87	80.70	82.05	80.46
type of high school:	lyceum	0.838	0.832	0.789	0.832	0.795
	others	0.162	0.168	0.211	0.168	0.205

Table 6: *Weighted means (or proportions) for each pre-treatment covariate included in the multinomial logit model used to compute individual weights.*

	Latent state (h)			
Contract type ($j = 1$)	1	2	3	4
none	1.000	0.000	0.000	0.000
temporary	0.000	0.964	0.994	0.000
permanent	0.000	0.036	0.006	1.000

	Latent state (h)			
Skill ($j = 2$)	1	2	3	4
none	1.000	0.000	0.000	0.000
low/medium	0.000	0.991	0.000	0.302
high	0.000	0.009	1.000	0.698

	Latent state (h)			
Gross income ($j = 3$)	1	2	3	4
none	1.000	0.014	0.033	0.020
$\leqslant 3750$	0.000	0.707	0.650	0.256
>3750	0.000	0.279	0.317	0.724

Table 7: *Estimated conditional probabilities of labor condition ($\phi_{jy|h}$), under the proposed model for the graduates in Milan with $k = 4$ latent states.*

	Latent state (h)		
Treatment	2	3	4
technical ($\hat{\alpha}_h$)	-1.239**	-0.584**	-1.005**
architecture vs. technical ($\hat{\beta}_{h2}$)	-1.177**	-1.267**	-1.369**
economic vs. technical ($\hat{\beta}_{h3}$)	0.405	-0.232	-0.118
humanities vs. technical ($\hat{\beta}_{h4}$)	-0.522	-0.776**	-1.623**
scientific vs. technical ($\hat{\beta}_{h5}$)	-0.434	-0.981†	-0.903
economic vs. architecture ($\hat{\beta}_{h3} - \hat{\beta}_{h2}$)	1.582**	1.036**	1.251*
humanities vs. architecture ($\hat{\beta}_{h4} - \hat{\beta}_{h2}$)	0.655†	0.492	-0.254
scientific vs. architecture ($\hat{\beta}_{h5} - \hat{\beta}_{h2}$)	0.743	0.287	0.466
humanities vs. economic ($\hat{\beta}_{h4} - \hat{\beta}_{h3}$)	-0.927**	-0.544*	-1.504**
scientific vs. economic ($\hat{\beta}_{h5} - \hat{\beta}_{h3}$)	-0.839	-0.749	0.784
scientific vs. humanities ($\hat{\beta}_{h5} - \hat{\beta}_{h4}$)	0.088	-0.205	0.720

Table 8: *Estimates of the logit regression parameters affecting the initial probabilities of the latent process under the selected LM causal model with $k = 4$ latent states (†significant at 10%, *significant at 5%, **significant at 1%).*

	Latent state (h)			
Treatment	1	2	3	4
technical	0.452	0.131	0.252	0.165
architecture	0.747	0.067	0.116	0.070
economic	0.454	0.197	0.201	0.148
humanities	0.666	0.115	0.171	0.048
scientific	0.647	0.121	0.136	0.096

Table 9: *Estimated initial probabilities for each type of treatment under the selected causal LM model with $k = 4$ latent states.*

	Latent state (h)		
Treatment	2	3	4
technical $\bar{h} = 1$ ($\hat{\gamma}_{1h}$)	-3.020**	-1.620**	-1.894**
technical $\bar{h} = 2$ ($\hat{\gamma}_{2h}$)	1.880**	-0.226	0.405
technical $\bar{h} = 3$ ($\hat{\gamma}_{3h}$)	-2.497†	2.185**	0.171
technical $\bar{h} = 4$ ($\hat{\gamma}_{4h}$)	-14.382**	-0.039	5.934**
architecture vs. technical ($\hat{\delta}_{h2}$)	-0.368	-0.862**	-2.639**
economic vs. technical ($\hat{\delta}_{h3}$)	0.080	-0.653*	-1.275**
humanities vs. technical ($\hat{\delta}_{h4}$)	-0.172	-0.524*	-1.851**
scientific vs. technical ($\hat{\delta}_{h5}$)	-0.179	-0.769*	-1.507**
economic vs. architecture ($\hat{\delta}_{h3} - \hat{\delta}_{h2}$)	0.448	0.208	1.364**
humanities vs. architecture ($\hat{\delta}_{h4} - \hat{\delta}_{h2}$)	0.196	0.338	0.788†
scientific vs. architecture ($\hat{\delta}_{h5} - \hat{\delta}_{h2}$)	0.189	0.092	1.132†
humanities vs. economic ($\hat{\delta}_{h4} - \hat{\delta}_{h3}$)	-0.252	0.130	-0.576
scientific vs. economic ($\hat{\delta}_{h5} - \hat{\delta}_{h3}$)	-0.259	-0.116	-0.232
scientific vs. humanities ($\hat{\delta}_{h5} - \hat{\delta}_{h4}$)	-0.007	-0.246	0.344

Table 10: *Estimates of the logit regression parameters affecting the transition probabilities of the latent process under the selected causal LM model (†significant at 10%, *significant at 5%, **significant at 1%).*

		Latent state (h)			
Degree	$\bar{h}$	1	2	3	4
technical	1	0.716	0.035	0.142	0.107
	2	0.102	0.665	0.081	0.152
	3	0.009	0.007	0.797	0.106
	4	0.003	0.000	0.002	0.995
architecture	1	0.886	0.030	0.074	0.010
	2	0.167	0.758	0.057	0.018
	3	0.204	0.012	0.767	0.017
	4	0.036	0.000	0.014	0.950
economic	1	0.835	0.044	0.086	0.035
	2	0.112	0.795	0.046	0.047
	3	0.165	0.015	0.765	0.055
	4	0.009	0.000	0.005	0.986
humanities	1	0.846	0.035	0.099	0.020
	2	0.138	0.764	0.065	0.033
	3	0.152	0.011	0.808	0.029
	4	0.016	0.000	0.009	0.974
scientific	1	0.858	0.034	0.079	0.029
	2	0.139	0.764	0.051	0.046
	3	0.183	0.013	0.756	0.048
	4	0.012	0.000	0.005	0.983

Table 11: *Estimates of the transition probabilities under the selected causal LM model referred to each treatment.*