Identification and estimation of average causal effects when treatment status is ignorable within unobserved strata

John Gardner*

Abstract

This paper extends nonparametric matching and propensity-score reweighting methods to settings in which unobserved variables influence both treatment assignment and counterfactual outcomes. Identification proceeds under the assumption that counterfactual outcomes are independent of treatment status conditional on observed covariates and membership in one of a finite set of latent classes. Individuals are first assigned to latent classes according to posterior probabilities of class membership derived from a finite-mixture model that relates a set of auxiliary variables to latent-class membership. Average causal effects are then identified by comparing outcomes among treated and untreated individuals assigned to the same class, correcting for misclassifications arising in the first step. The identification procedure suggests computationally attractive latent-class matching and propensity-score reweighting estimators that obviate the need to directly estimate the distributions of counterfactual outcomes. In Monte Carlo studies, the resulting estimates are centered around the correct average causal effects with minimal loss of precision compared to competing estimators that misstate those effects.

Keywords: Treatment effects, causal effects, endogeneity, unobserved heterogeneity, finite mixtures, matching, propensity-score reweighting.

JEL Codes: C21, C22.

1 Introduction

Nonparametric treatment effect estimators such as matching and inverse-probability weighting are popular, at least in part, because they take a simple and intuitive approach to

*Department of Economics, University of Mississippi. Email: jrgardne@olemiss.edu. I thank John Conlon, Natalia Kolesnikova and Alice Sheehan for helpful comments.

identifying causal effects that does not rely on restrictive functional form or distributional assumptions (Imbens and Wooldridge, 2009; Imbens, 2014, provide excellent reviews). Despite this elegance, the applicability of these estimators is often limited by their predication on the assumption of selection on observables (equivalently, conditional independence, unconfoundedness, or exogeneity), which asserts that, conditional on a set of observed covariates, the counterfactual outcomes that individuals would experience with and without the treatment are independent of whether they actually received the treatment. In observational studies, there is rarely reason to believe that all of the variables that are simultaneously related to both counterfactual outcomes and the treatment decision are observed in the data, raising the possibility that comparisons of outcomes between treated and untreated units are contaminated with bias due to self-selection, even conditional on observed covariates, and do not represent the causal effect of the treatment on outcomes.

Heckman et al. (1997, 1998) and Abadie (2005) develop difference-in-differences matching and propensity-score reweighting methods that use panel variation in outcomes to relax the requirement of selection on observables. These methods identify average causal effects under a nonparametric version of the parallel trends assumption used in traditional difference-in-differences research designs: conditional on observed covariates, the changes in outcomes that individuals would experience absent the treatment are independent of treatment status. Under this assumption, the bias due to selection on unobservables is the same both before and after the treatment is made available, and can therefore be eliminated by subtracting pre-treatment differences between treated and untreated units from those same differences recorded in a post-treatment period.[1]

This paper develops a different approach to identifying average causal effects when unobserved variables affect both treatment status and counterfactual outcomes. Instead of eliminating biases introduced by failure to condition on such unobservables, however, the methods developed below exploit panel variation in observed variables in order to identify average counterfactual outcomes and causal effects conditional on both the observed and unobserved determinants of treatment status and counterfactual outcomes. This approach proceeds from the assumptions that each individual belongs to one of a finite set of latent classes or unobserved types and that treatment status is ignorable—that is, counterfactual outcomes are independent of treatment status and the probability of receiving the treatment

[1] In addition, Abadie and Gardeazabal (2003) and Abadie et al. (2010) develop synthetic control methods that use a weighted average of outcomes for untreated units to approximate the outcomes that the treated unit would have experienced absent the treatment, showing that the approximation error disappears as the number of pre-treatment periods used to form the synthetic controls increases. While this method is closely related to those developed in Heckman et al. (1997, 1998) and Abadie (2005), it applies primarily to settings where only one observational unit receives the treatment, while the difference-in-differences methods discussed above apply when there are multiple treated and untreated units.

is strictly between zero and one (Rosenbaum and Rubin, 1983)—conditional on this latent class and a set of observed covariates.

From these assumptions, identification of average causal effects proceeds in two steps. In the first step, a finite-mixture model is used to recover the likelihood of a set of auxiliary dependent variables and covariates as a sum of latent-class-specific likelihoods, weighted by the prior (that is, unconditional) probabilities of membership in each latent class. The purpose of this part of the procedure is to assign individuals to latent classes (that is, impute the classes to which they belong) on the basis of their posterior probabilities of class membership, given their realizations of the observed auxiliary variables. For example, the first step of the procedure might consist of a model of the history of pre-treatment-period outcomes as a function of latent-class membership and time-invariant covariates, or if the treatment decision is made repeatedly, a dynamic discrete-choice model of the treatment decision as a function of latent-class membership and (potentially time-varying) covariates.

In the second step, observed-covariate×latent-class-specific average counterfactual outcomes are identified by computing average outcomes conditional on observed covariates and assigned latent classes, then correcting for potential errors in the latent-class assignments. These average counterfactual outcomes can then be used to identify covariate×latent-class-specific average causal effects which, in turn, can be aggregated to the population and treated-population levels to identify the average effect of the treatment (ATE) and average effect of the treatment on the treated (ATT).

Though the conditions under which this approach can be used to identify average causal effects are broadly similar to those for the non- and semi-parametric difference-in-differences methods described above, the latent-class approach has several advantages. Unlike difference-in-differences methods, it places no restrictions on the influence of latent variables on outcomes. For example, it can be applied when the effect of those variables on outcomes changes over time—a possibility that would invalidate the parallel-trends assumption. It can also be applied when outcomes are observed only at a single point in time, provided that there are sufficient non-outcome auxiliary variables with which to identify the first-step finite-mixture model. In addition, rather than remove the bias due to selection on unobservables, it identifies average causal effects conditional on those unobservables; these latent-class-specific effects may be of intrinsic interest.

A drawback of the latent-class approach is that it requires specification and identification of an auxiliary finite mixture model. In many cases, finite-mixture models are nonparametrically identified from variation in observed variables, intuitively because successive realizations of these variables reveal information about latent classes to which individuals likely belong. However, identification requires restrictions on the number and dimension of the observed

variables included in the model and on the latent structure through which these variables are related; I discuss specification and identification of the first-step finite-mixture model in greater detail below.

This is not the first paper to combine finite-mixture models with matching and propensity-score methods. This paper's closest progenitors are series of papers by Haviland and Nagin (2005), Haviland et al. (2008), and Bartolucci et al. (2012a,b). Haviland and Nagin (2005) and Haviland et al. (2008) use latent-class assignments based on a finite-mixture model of pre-treatment-period delinquent behavior as a device for matching gang members to non-members with comparable histories of such behavior. They use within-assigned-class comparisons of members and non-members in order to estimate the effect of gang membership on delinquency under the assumption that delinquent behavior is independent of gang membership conditional on observed behavioral histories. The method proposed by Bartolucci et al. (2012a,b) uses a finite-mixture model of the evolution of the treatment decision and observed covariates to obtain latent-class-specific propensity scores. The second step of their procedure computes treatment effects using propensity-score reweighting methods within assigned classes under the assumption that counterfactual outcomes are independent of treatment status given observed covariates and latent-class membership.

As I discuss below, the methods used in these papers do not identify average causal effects within observed-covariate$\times$latent-class strata. The reason for this is that the latent classes to which individuals are assigned may not be the classes to which they actually belong. In other words, latent-class assignments based on posterior probabilities of class membership obtained from an auxiliary finite-mixture model are error-ridden measures of true class membership. The procedure developed below uses information recovered from the auxiliary finite-mixture model to identify the distributions of, and ultimately correct for, these classification errors. This approach draws on work by Bolck et al. (2004), who noted a similar misclassification problem in the literature on regressions of latent-class assignments on external covariates (also see Vermunt, 2010; Bray et al., 2015).

An advantage of this misclassification-correction approach is that it does not require that outcomes are modeled directly in the finite-mixture model used in the first step of the procedure. This reduces the dimension of the identification problem, making the procedure computationally attractive to implement. Another solution to the misclassification problem is to include observed outcomes and treatment status among the variables modeled in the first step of the procedure in order to directly identify the distributions of counterfactual outcomes conditional on covariates and latent-class membership.[2] Though the misclassification-correction approach is the focus of this paper, I discuss the direct approach

[2]Similar solutions have been noted in the literature on latent-class regression (see Bolck et al., 2004).

in Appendix A.

I develop the identification and estimation procedure as follows. In Section 2, I discuss the assumptions and data requirements of the procedure. In Section 3, I discuss identification, specification, and estimation of the first-step finite-mixture model, illustrating with examples from the literature. I begin Section 4 by showing how average causal effects are identified in a second step from observed outcomes and error-ridden latent-class assignments. I then develop latent-class matching and reweighting estimators motivated by the identification results. In Section 5, I implement Monto-Carlo studies that illustrate the small-sample performance of the estimators. I conclude in Section 6.

2 Preliminaries

Let D_{it} be an indicator for whether individual i receives a (binary) treatment at time t. Let Y_{dit} denote the possibly counterfactual outcome that i would experience if assigned to treatment status $d \in \{0, 1\}$ at time t, so that the time-t causal effect of the treatment on unit i can be expressed as $Y_{1it} - Y_{0it}$ and realized outcomes can be expressed as $Y_{it} = (1 - D_{it})Y_{0it} + D_{it}Y_{1it}$. The methods developed below can be applied in settings where the treatment decision is made repeatedly (in which case the treatment effect may depend on time and can be estimated at each of the $t \in \{1, \ldots, T\}$ periods in which it is available) as well as in settings where the treatment decision is permanent and made only once.

The approach developed below identifies average causal effects under the assumptions that each individual is characterized by time-invariant membership J_i in one of a finite set of latent classes and that, at time t, treatment status is strongly ignorable conditional on latent-class membership J_i and a set X_{it} of observed covariates. Following Rosenbaum and Rubin (1983), strong ignorability in this context means that counterfactual outcomes are independent of treatment status conditional on observed covariates and latent-class membership, and that treated and untreated individuals have overlapping characteristics in the sense that the probability of receiving the treatment is strictly between zero and one conditional on covariates and latent-class membership. Formally:

Assumption 1. *Counterfactual outcomes and treatment assignment satisfy*

$$(Y_{0it}, Y_{1it}) \perp\!\!\!\perp D_{it} | X_{it}, J_i, \tag{1}$$

and

$$P(D_{it} = 1 | X_{it}, J_i) \in (0, 1), \tag{2}$$

where $J_i \in \{1, \ldots, |J|\}$.

Throughout the remainder of this paper, I assume for simplicity that the observed covariates X_{it} are either discrete or have been discretized. Also note that while, in contrast to selection-on-observables research designs, the latent-class overlap component (2) of Assumption 1 is not directly verifiable, it can be assessed as part of the identification procedure developed below.

The environment established in Assumption 1 is similar to that required for causal inference under the difference-in-differences methods described above, which can be motivated by a model in which outcomes depend on unobserved, time-invariant, and additively separable fixed effects. However, Assumption 1 places no restrictions on the relationship between outcomes and their unobserved determinants J_i. While the assumption that those unobserved determinants are drawn from a discrete distribution is restrictive, the number of latent classes with which average causal effects can be identified is increasing in the dimension (in a sense made precise below) of the observed auxiliary variables. Consequently, the methods developed below can be viewed as approximations that improve as more data are gathered.[3]

The identification procedure requires that the data contain observations on a set of discrete (or discretized) auxiliary dependent variables Z_{is} and auxiliary covariates W_{is}, which I index by $s \in \{1, \ldots, S\}$ to allow for the possibility that they are recorded at different times than the variables (Y_{it}, D_{it}, X_{it}) included in the causal model. The procedure also requires that treatment status D_{it} and the observed covariates X_{it} on which Assumption 1 depends are included in the first-step finite-mixture model, and that the remaining variables modeled in this step are independent of counterfactual outcomes given X_{it} and J_i:

Assumption 2. *X_{it} and D_{it} are elements of (Z_{is}, W_{is}) for some $s \in \{1, \ldots, S\}$, and*

$$(Y_{0it}, Y_{1it}) \perp\!\!\!\perp Z_{is} | X_{it}, J_i \tag{3}$$

for all $s \in \{1, \ldots, S\}$.

Assumption 2 restricts the models used to identify average casual effects, not the data-generating processes themselves. For example, if there is reason to believe that counterfactual outcomes depend on some element of the vector Z_{is}, then this element should be included in X_{it} (or excluded from Z_{is}). Note that the identification procedure does not require observed outcomes Y_{it} to be included in the auxiliary model.

[3] In this sense, the latent-class approach is similar to the synthetic control methods developed in Abadie (2005) and Abadie et al. (2010).

3 Step one: The auxiliary model

The first step of the identification procedure involves using an auxiliary finite-mixture model in order to recover individuals' posterior probabilities of class membership, conditional on their realizations of the variables included in the model. Perhaps because finite-mixture models have traditionally been used in parametric settings in which their identification has been somewhat mysterious, they have not yet found widespread use in econometrics (with some notable exceptions, see Heckman and Singer, 1984a,b). However, as a growing literature has shown that these models are often identified, without distributional assumptions or parametric restrictions, from variation in observed variables (see Hall and Zhou, 2003; Allman et al., 2009; Kasahara and Shimotsu, 2009; Hu and Shum, 2012; Henry et al., 2014; Bonhomme et al., 2016; Compiani and Kitamura, 2016), they have found increasing application in settings with unobserved heterogeneity (see, for example Arcidiacono and Jones, 2003; Arcidiacono and Miller, 2011; Aguirregabiria and Mira, 2016).

3.1 Specification and identification

Dropping the i subscripts from random variables for simplicity, let $Z = (Z_1, \ldots, Z_S)$ and $W = (W_1, \ldots, W_S)$ denote the sequences of auxiliary dependent variables and covariates. When each individual belongs to a latent class $J \in \{1, \ldots, |J|\}$, the likelihood of observing the sequence (z, w) of auxiliary variables conditional on an initial realization w_1 of the auxiliary covariates can be written

$$\begin{aligned}\ell(z, w|w_1) &= \sum_{j=1}^{|J|} P(J = j|w_1)\ell(z, w|j, w_1) \\ &= \sum_{j=1}^{|J|} P(J = j|w_1) \prod_{s=2}^{S} P(Z_s = z_s|j, w_s, \ldots, w_1, z_{s-1}, \ldots, z_1) \\ &\qquad \times P(W_s = w_s|j, z_{s-1}, \ldots, z_1, w_{s-1}, \ldots, w_1) P(Z_1 = z_1|j, w_1).\end{aligned} \tag{4}$$

Nonparametric identification in this context refers to recovery of the conditional distributions of the observed variables and latent classes that enter into the likelihood function. For the moment, assume that the number $|J|$ of latent classes is known (I discuss identification and estimation of this number below). Though the precise minimal requirements for nonparametric identification of finite-mixtures taking the general form in (4) are not known, identification has been established for a number of important cases. Which, or whether any, case is relevant depends on prior hypotheses about the underlying model.

In the first of these cases, the auxiliary dependent variables Z_s, $s \in \{1, \ldots, S\}$, are

independent of one another conditional on a set of time-invariant auxiliary covariates W and latent-class membership J. In this case, (4) becomes

$$\ell(z|w) = \sum_{j=1}^{|J|} P(J = j|w) \prod_{s=1}^{S} P(Z_s = z_s|j, w). \tag{5}$$

When the auxiliary variables are discrete, finite-mixtures of this form are known simply as latent-class models. Allman et al. (2009) show that such models are nonparametrically identified for any number $|J|$ of latent classes, provided that enough auxiliary variables Z_s are observed or that the support of the observed variables is of sufficient dimension. For example, if each of the Z_s have k points of support, they show that a sufficient condition for identification is that $S \geq 2\lceil \log_k |J| \rceil + 1$ (where $\lceil \cdot \rceil$ is the integer ceiling function, see Allman et al. 2009, Section 5).[4]

Such models are appropriate when latent-class membership and observed covariates are the source of serial correlation between the auxiliary variables modeled in the first step. The following example illustrates how these models can be used as the first in a two-step procedure for identifying average causal effects (the Monte Carlo studies presented in Section 5 below provide another example).

Example 1. Haviland and Nagin (2005) and Haviland et al. (2008) estimate the causal effects of gang membership on delinquent behavior under the assumption that counterfactual behavior at time t is independent of gang membership conditional on histories $Z_1, \ldots, Z_S$ of such behavior observed in a pre-treatment period (that is, before anyone has joined a gang). Instead of comparing gang members and non-members with similar behavioral histories, they classify individuals into latent classes (known as trajectory groups) using a finite-mixture model in which successive observations of pre-treatment-period behavior are independent of each other conditional on trajectory-group membership. They show that the distributions of observed behavioral histories are comparable within trajectory groups, then estimate the effect of gang membership by comparing behavioral outcomes among members and non-members assigned to the same trajectory groups.

While these papers use assigned latent-class membership as a device for finding *observably* similar gang members and non-members, a natural extension of the idea is to identify the causal effects of gang membership under the assumption that counterfactual behavioral outcomes are independent of gang membership conditional on *unobserved* latent-class membership J. This could be accomplished using a first-step finite-mixture model of pre-

[4]Strictly speaking, in the discrete case Allman et al. (2009) provide conditions for generic identifiability, meaning that the set of latent-class-specific mass functions on which identification fails is of Lebesgue measure zero. They also establish identification for the case of continuous Z, subject to linear independence conditions.

treatment-period behavioral histories in which successive observations of delinquent behavior are mutually independent within latent classes (and conditional on treatment status):

$$\ell(z_1, \ldots, z_S|d) = \sum_{j=1}^{|J|} P(J = j|D = d) \prod_{s=1}^{S} P(Z_s = z_s|j, d). \tag{6}$$

As I discuss below, this first-step model could then be used to assign individuals to latent classes in order to identify latent-class-specific average causal effects. The two-step procedure for identifying average causal effects does not require that behavior at time t is included in the finite-mixture model, and can be applied even when the first step uses coarsely discretized measures of pre-treatment-period behavior (e.g., by binning continuous behavioral outcomes into quartiles). In addition, the finite-mixture model can be identified within observed-covariate strata to allow for the possibility that treatment status depends on these covariates as well as latent-class membership. Finally, note that (6) does not require that the conditional distributions of the auxiliary variables are time invariant.

Nonparametric identification is also well understood when the auxiliary dependent variables and covariates exhibit dynamic dependence that follows a Markov structure, as in standard discrete-choice dynamic programming models, conditional on latent-class membership. In the first-order Markov case, (4) can be written

$$\begin{aligned} \ell(z, w|w_1) = \sum_{j=1}^{|J|} P(J = j|w_1) \prod_{s=2}^{S} P(Z_s = z_s|j, w_s) P(W_s = w_s|j, w_{s-1}, z_{s-1}) \\ \times P(Z_1 = z_1|j, w_1). \end{aligned} \tag{7}$$

Kasahara and Shimotsu (2009, also see Hu and Shum 2012) establish conditions under which several such models are nonparametrically identified. The identification results, which require restrictions on the number of time periods observed and rank conditions that relate to the number of unobserved types, differ depending on whether the conditional distributions of the auxiliary variables are stationary, whether lagged dependent variables affect the covariate transitions, and whether the covariate transitions depend on latent-class membership.

Another example from the literature illustrates how finite-mixture models like (7), in which the auxiliary variables are related through relatively complicated dynamic structures, can be incorporated into the two-step identification procedure.

Example 2. Bartolucci et al. (2012a,b) apply their two-step procedure (described in the introduction) to estimate the effect of wage subsidies to firms on employment, allowing for the possibility that unobserved firm characteristics affect both the receipt of wage subsidies

and counterfactual employment levels. In the first step of their procedure, they use a finite-mixture model of the joint evolution of wage subsidies and observed covariates to assign individuals to latent-classes. They model receipt of wage subsidies (Z_s) using finite-mixture logit models that depend on latent-class membership J and first- and second-order lagged values (W_s) of receipt of subsidies, employment, wages, capital measures, sales, profits, and time effects. They model the evolution these covariates using finite mixtures of parametric multivariate regressions of the covariates on their lagged values.

This is a non-stationary second-order-Markov finite-mixture model in which the transitions between covariates depend on both latent-class membership and lags of the dependent variable. As Kasahara and Shimotsu (2009) note, whether such models are identified without parametric assumptions is not known.[5] However, the results in Kasahara and Shimotsu (2009, Proposition 6 and following remarks) show that a version of this model in which the conditional distributions of the auxiliary variables do not depend on time, and the transitions between auxiliary covariates are independent of latent-class membership, is nonparametrically identified under mild regularity conditions from observations on nine periods of data (and six periods when the Markov structure is first order).

3.2 Estimation

In principle the auxiliary model can be estimated nonparametrically by directly maximizing the sample analog of $E\{\log[\ell(Z, W|W_1; \theta)]\}$ with respect to the vector θ of unconstrained components of the likelihood function (that is, the conditional distributions of the observed and latent variables that comprise it). Alternatively, these components can be specified as smooth functions of a vector of parameters in order to implement the identification procedure semi-parametrically. Because finite-mixture log likelihoods can be difficult to maximize directly, they are often estimated using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977, see Arcidiacono and Miller 2011 for a discussion of this approach in a dynamic discrete-choice setting).[6] The Monte Carlo studies presented below provide an example of this approach. Many statistical packages include some facility for estimating finite-mixture models.[7]

Another important consideration in estimating finite mixtures is the number of latent

[5]Identification under parametric distributional assumptions follows more readily (see Teicher, 1963; Grün and Leisch, 2008b).

[6]Wu (1983) showed that the EM algorithm may converge to flat points of the log-likelihood function that are not global maxima. A typical solution is to initialize the algorithm at a number of different starting values and choose the solution with the highest log likelihood (also see Arcidiacono and Jones, 2003).

[7]Kasahara and Shimotsu (2009) and Bonhomme et al. (2016) develop alternative nonparametric estimators of finite mixtures that can be used even when the auxiliariy variables are continuous.

classes that should be included in the model. The most common approach is to estimate multiple models, each with a different number of latent classes, and choose the model that maximizes either the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) (see, e.g., Grün and Leisch, 2008a). In addition, Kasahara and Shimotsu (2009, 2014) show that the number of classes, or a bound on that number, can be identified and estimated nonparametrically.

4 Step two: Recovering average causal effects

Denote by q_{ji} the posterior probability that individual i belongs to class $j \in \{1, \ldots, |J|\}$, given their realizations of the auxiliary variables (Z_i, W_i). These posterior probabilities can be expressed in terms of latent-class-specific likelihoods using Bayes' rule as

$$q_{ji} \equiv P(J_i = j|z_i, w_i) = \frac{\ell(z_i, w_i|j, w_{1i})}{\sum_{k=1}^{|J|} P(J_i = k|w_{1i})\ell(z_i, w_i|k, w_{1i})},$$

and are therefore identified along with the auxiliary model.[8]

Two-step analyses based on finite-mixture models typically proceed by assigning individuals to latent classes according to posterior probabilities of class membership recovered from an auxiliary model, then examining the relationships of interest within assigned classes. In such studies, two procedures for assigning individuals to latent classes are common. Under modal assignment (also known as hard assignment or the classify-analyze method), individuals are assigned to the latent class $\hat{J}_i$ for which their posterior probability of membership is greatest:

$$\hat{J}_i = \underset{j \in \{1,\ldots,|J|\}}{\text{argmax}}\ q_{ji}.$$

Under proportional assignment (also known as soft assignment or the expected-value method), individuals are assigned to each latent class in proportion to their posterior probabilities of membership in that class.

Haviland and Nagin (2005), Haviland et al. (2008), and Bartolucci et al. (2012a,b) use these procedures to assign individuals to latent classes (or observed-covariate×latent-class strata), then estimate treatment effects by comparing outcomes between treated and untreated individuals assigned to the same classes. As I note in the introduction, because the classes to which individuals are assigned under these procedures may not be the classes to which they actually belong, such comparisons do not identify latent-class-specific average

[8]Most pre-programmed routines for estimating finite-mixture models include estimates of these posteriors as part of their output.

causal effects.

As the following results show, the posterior probabilities of latent-class membership can be used to recover the distributions of, and ultimately correct for, these classification errors, in order to identify average counterfactual outcomes and causal effects within observed-covariate×latent-class strata. This insight is originally due to Bolck et al. (2004), who studied a related misclassification problem in the literature on latent-class regressions (also see Vermunt, 2010; Lanza et al., 2013; Bray et al., 2015).[9]

Another solution to this problem is to include time-t observed outcomes Y_t among the variables (Z, W) modeled in the first step of the identification procedure.[10] This approach works by directly identifying the distributions of counterfactual outcomes within covariate×latent-class strata (which, under Assumption 1, can be obtained from the distributions of observed outcomes conditional on treatment status, covariates, and latent-class membership). The advantage of the indirect misclassification-correction approach is that it obviates the need to identify the distributions of counterfactual outcomes. This reduces the dimension of the identification problem in the first-step of the procedure (particularly when outcomes are continuous or discrete with high-dimensional support), making the resulting estimators simple, and computationally attractive, to implement. I discuss the direct approach further in Appendix A.

4.1 Modal assignment

The average counterfactual outcome that an individual with covariates $X_t = x_t$ belonging to latent-class $J = j$ would experience at time t if assigned treatment status $D_t = d_t$ is $E(Y_{dt}|x_{dt}, j)$. Modal assignment methods impute these counterfactual outcomes by averaging observed outcomes among those with covariates x_t, treatment status d_t, and assigned to latent class $\hat{J} = j$:

$$E(Y_t|\hat{J} = j, d_t, x_t). \tag{8}$$

Proposition 1, below, shows that (8) identifies a within-covariate sum of latent-class-specific average counterfactual outcomes, weighted by the probabilities that members of each class $k \in \{1, \ldots, |J|\}$ are modally assigned to class j. The proposition also shows how these classification errors can be corrected in order to identify average causal effects within covariate×latent-class strata. Because the proof is simple and instructive, I present it in the main text.

[9] In that literature, the object of analysis is the effect of observed covariates on latent-class membership.

[10] Similar solutions have been noted in the latent-class regression literature (see Bolck et al., 2004; Vermunt, 2010; Bray et al., 2015).

Proposition 1. *Let $E_{Y_t|\hat{J},d_t,x_t}$ be the $|J|$-vector with jth element $E(Y_t|\hat{J}=j,D_t=d_t,X_t=x_t)$, $E_{Y_{dt}|J,x_t}$ be the $|J|$-vector of average counterfactual outcomes with jth element $E(Y_{dt}|J=j,X_t=x_t)$, and P_{d_t,x_t} be the $|J|\times|J|$ matrix with (j,k)th element $P(J=k|\hat{J}=j,D_t=d_t,X_t=x_t)$. Under Assumptions 1 and 2,*

$$E_{Y_{dt}|J,x_t} = P_{d_t,x_t}^{-1} E_{Y_t|\hat{J},d_t,x_t}$$

for any (d_t,x_t) such that P_{d_t,x_t} is invertible. Furthermore, the (j,k)th element of P_{d_t,x_t} satisfies

$$P(J=k|\hat{J}=j,d_t,x_t) = E\left(\frac{q_k 1(\hat{J}=j)}{P(\hat{J}=j|d_t,x_t)}\middle| D_t=d_t,X_t=x_t\right).$$

Proof. Modal assignment methods impute counterfactual outcomes using (8), which can be expressed via the law of total expectation as

$$E(Y_t|D_t=d_t,X_t=x_t,\hat{J}=j) = \sum_{k=1}^{|J|} E(Y_t|d_t,x_t,\hat{J}=j,J=k)P(J=k|d_t,x_t,\hat{J}=j). \quad (9)$$

Conditional on $D_t=d_t$, observed outcomes Y_t can be replaced with counterfactual outcomes Y_{dt} in the right-hand side of (9). Because counterfactual outcomes are independent of treatment status conditional on X_t and J under Assumption 1, the condition that $D_t=d_t$ can be dropped from the right-hand-side expectations. Furthermore, because counterfactual outcomes are also independent of (Z,W) conditional on X_t and J under Assumption 2, the condition that $\hat{J}=j$, which is a function of (Z,W), can also be excluded from these expectations. This shows that the modal assignment method identifies a weighted average of latent-class specific causal effects:

$$E(Y_t|D_t=d_t,X_t=x_t,\hat{J}=j) = \sum_{k=1}^{|J|} E(Y_{dt}|x_t,J=k)P(J=k|d_t,x_t,\hat{J}=j). \quad (10)$$

Stacking (10) for different values of $j\in\{1,\ldots,|J|\}$ gives

$$E_{Y_t|\hat{J},d_t,x_t} = P_{d_t,x_t} E_{Y_{dt}|J,x_t},$$

which proves the first part of the proposition.

To prove the second part, note that the probability that an observation belongs to class k conditional on covariates X_t, treatment status D_t, and modal assignment into class j, can

be expressed as

$$
\begin{aligned}
P(J=k|x_t,d_t,\hat{J}=j) &= \frac{P(J=k,\hat{J}=j|d_t,x_t)}{P(\hat{J}=j|d_t,x_t)} \\
&= \frac{\sum_{z,w} P(J=k|\hat{J}=j,z,w)P(\hat{J}=j|z,w)P(z,w|d_t,x_t)}{P(\hat{J}=j|d_t,x_t)} \\
&= \frac{\sum_{z,w} P(J=k|z,w)1(\hat{J}=j)P(z,w|d_t,x_t)}{P(\hat{J}=j|d_t,x_t)} \\
&= E\left(\frac{q_k 1(\hat{J}=j)}{P(\hat{J}=j|d_t,x_t)}\middle| d_t,x_t\right),
\end{aligned}
\tag{11}
$$

where $1(\cdot)$ is the indicator function and the sums run over the support $\text{supp}(Z,W)$ of the joint distribution of Z and W. The second equality in (11) follows because (D_t, X_t) is an element of (Z,W) under Assumption 2. The third follows because $\hat{J}$ is a deterministic function of (Z,W). The fourth follows from the definition of q_k. □

4.2 Proportional assignment

Under proportional assignment, mean counterfactual outcomes within (X_t, J) strata are imputed as

$$
E\left(\frac{Y_t q_j}{E(q_j|x_t,d_t)}\middle| x_t,d_t\right) = E\left(\frac{Y_t q_j}{P(J=j|x_t,d_t)}\middle| x_t,d_t\right). \tag{12}
$$

Proposition 2 shows that, like its modal-assignment counterpart (8), (12) also identifies within-covariate-strata weighted sums of latent-class-specific average causal effects, from which the average causal effects themselves can be recovered by correcting for misclassification.

Proposition 2. *Let $E_{Y_t q_J|d_t,x_t}$ be the $|J|$-vector with jth element $E(Y_t q_j|D_t=d_t,X_t=x_t)/E(q_j|D_t=d_t,X_t=x_t)$, $E_{Y_{dt}|J,x_t}$ be the $|J|$-vector of average counterfactual outcomes defined as in Proposition 1, and Q_{d_t,x_t} be the $|J|\times|J|$ matrix with (j,k)th element $E[P(J=k|Z)|J=j,D_t=d_t,X_t=x_t]$. Under Assumptions 1 and 2,*

$$
E_{Y_{dt}|J,x_t} = Q^{-1}_{d_t,x_t} E_{Y_t q_J|d_t,x_t}
$$

for any (d_t,x_t) such that Q_{d_t,x_t} is invertible. Furthermore, the (j,k)th element of Q_{d_t,x_t} satisfies

$$
E[P(J=k|Z)|J=j,D_t=d_t,X_t=x_t] = E\left(\frac{q_j q_k}{P(J=j|d_t,x_t)}\middle| D_t=d_t,X_t=x_t\right).
$$

Proof. Under Assumptions 1 and 2, (12) can be written

$$
\begin{aligned}
E\left(\frac{Y_t q_j}{P(J=j|d_t,x_t)}\middle|d_t,x_t\right) &= \sum_{z,w} E\left(\frac{Y_t q_j}{P(J=j|d_t,x_t)}\middle|z,w,d_t,x_t,\right) P(z,w|d_t,x_t) \\
&= \sum_{z,w}\left(\sum_k \frac{q_j E(Y_{dt}|J=k,z,w,d_t,x_t)}{P(J=j|d_t,x_t)} P(J=k|z,w,d_t,x_t)\right) \\
&\quad \times P(z,w|d_t,x_t) \\
&= \sum_k \left(\sum_{z,w} \frac{q_j q_k P(z,w|d_t,x_t)}{P(J=j|d_t,x_t)}\right) E(Y_{dt}|J=k,x_t),
\end{aligned}
\tag{13}
$$

where $(z,w) \in \text{supp}(Z,W)$. The second equality in (13) uses the law of total expectation, the fact that observed outcomes Y_t can be replaced with counterfactual outcomes Y_{dt} conditional on treatment status $D_t = d_t$, and the fact that q_j is a function of (Z,W). The third equality follows from Assumptions 1 and 2, under which counterfactual outcomes are independent of D_t and (Z,W) conditional on X_t and J.

In addition, the expected proportional assignment of a member of class j into class k is

$$
\begin{aligned}
E[P(J=k|Z,W)|J=j,d_t,x_t] &= \sum_{z,w} E[P(J=k|z,w)|J=j,z,w,d_t,x_t] P(z,w|J=j,d_t,x_t) \\
&= \sum_{z,w} P(J=k|z,w) \frac{P(J=j|z,w,d_t,x_t)P(z,w|d_t,x_t)}{P(J=j|d_t,x_t)} \\
&= E\left(\frac{q_j q_k}{P(J=j|d_t,x_t)}\middle|d_t,x_t\right),
\end{aligned}
\tag{14}
$$

where I have used the facts that the q_j are functions of (Z,W) and that (D_t,X_t) is an element of (Z,W) under Assumption 2.

Together, (13) and (14) imply that $E_{Y_t q_J|d_t,x_t} = Q_{d_t,x_t} E_{Y_{dt}|J,x_t}$, and the conclusion follows. □

4.3 Population average causal effects and their reweighting interpretation

Propositions 1 and 2 show that the classification errors associated with both modal and proportional latent-class assignment can be corrected in order to use these procedures to identify average counterfactual outcomes within (X_t, J) strata. The differences between

average counterfactual treated and untreated outcomes can then be used to identify (X_t, J)-strata average causal effects as $ATE_t(x_t, j) = E(Y_{1t}|x_t, j) - E(Y_{0t}|x_t, j)$.[11] These strata-specific average causal effects, in turn, can be aggregated to the population and treated-population levels in order to identify the time-t average effect of the treatment (ATE) as

$$ATE_t = E(Y_{1t} - Y_{0t}) = \sum_{x_t,j} ATE_t(x_t, j)P(X_t = x_t, J = j), \tag{15}$$

and the average effect of the treatment on the treated (ATT) as

$$ATT_t = E(Y_{1t} - Y_{0t}|D_t = 1) = \sum_{x_t,j} ATE_t(x_t, j)P(X_t = x_t, J = j|D_t = 1), \tag{16}$$

where $(x_t, j) \in \text{supp}(X_t, J)$. The aggregation weights used in (15) and (16) are identified from the auxiliary model and observed covariates and treatment status as $P(x_t, j) = E(q_j|x_t)P(x_t)$ and $P(x_t, j|D_t = 1) = E(q_j|x_t, D_t = 1)P(x_t|D_t = 1)$.[12]

Many methods for identifying treatment effects under the assumption of selection on observables (including matching and difference-in-differences matching) can be interpreted as Horvitz-Thompson-style (1952) reweighting procedures that adjust for differences between the characteristics of the treated and untreated populations (see, e.g., Robins et al., 1994; Hirano et al., 2003; Abadie, 2005; Imbens and Wooldridge, 2009; Imbens, 2014). As Proposition 3 shows, a similar interpretation is available for the latent-class methods developed above.

Proposition 3. *Let $a_j \in \{1(\hat{J} = j), q_j\}$ be a procedure for determining assignment to latent class $j \in \{1, \ldots, |J|\}$, and let $A_{d_t,x_t} \in \{P_{d_t,x_t}, Q_{d_t,x_t}\}$ be the associated matrix of misclassification probabilities for each x_t and d_t in* $\text{supp}(X_t, D_t)$. *Under Assumptions 1 and 2, $ATE_t = E[Y_t(\omega_{1t} - \omega_{0t})]$ and $ATT_t = E[Y_t(\tilde{\omega}_{1t} - \tilde{\omega}_{0t})]$, where*

$$\omega_{dt} = \sum_j \sum_k A^{-1}_{d_t,X_t}(j, k) \frac{a_k 1(D_t = d_t)P(J = j|X_t)}{E(a_k|X_t, d_t)P(d_t|X_t)}$$

for $d_t \in \{0, 1\}$,

$$\tilde{\omega}_{1t} = \frac{1(D_t = 1)}{P(D_t = 1)}$$

[11] Note that under ignorability (Assumption 1), the average effect of the treatment and the average effect of the treatment on the treated are the same conditional on covariates and latent-class membership.

[12] Also note that if treatment occurs in multiple periods and the conditional distributions of the auxiliary variables are stationary, the time indices can be removed from (15) and (16) to identify treatment effects that are averaged over time (in addition to observed covariates and latent-class membership).

and

$$\tilde{\omega}_{0t} = \sum_j \sum_k A^{-1}_{1,X_t}(j,k) \frac{a_k 1(D_t = 0) P(J = j|X_t, D_t = 1) P(D_t = 1|X_t)}{E(a_k|D_t = 0, X_t) P(D_t = 0|X_t) P(D_t = 1)}.$$

The proof is presented in Appendix C. As noted above, the components of the weights ω_{dt} and $\tilde{\omega}_{dt}$ are identified from the auxiliary model and observed covariates and treatment status. To relate the latent-class reweighting procedures to those based on selection on observables (see Hirano et al., 2003), note that, were J observed, a_j would be an indicator for $J = j$ and A_{d_t,X_t} would be an identity matrix, so that $E(Y_t\omega_{dt})$ would become

$$\begin{aligned} E\left(\sum_j \frac{Y_t 1(J = j) 1(D_t = d_t)}{P(J = j|X_t, d_t)} \frac{P(J = j|X_t)}{P(D_t = d_t|X_t)}\right) &= E\left(\sum_j \frac{Y_t 1(J = j) 1(D_t = d_t) P(J = j|X_t)}{P(J = j, D_t = d_t|X_t)}\right) \\ &= E\left(\sum_j \frac{Y_t 1(J = j) 1(D_t = d_t)}{P(D_t = d_t|X_t, j)}\right) \\ &= E\left(\frac{Y_t 1(D_t = d_t)}{P(D_t = d_t|X_t, J)}\right), \end{aligned}$$

which is the population analog of the standard inverse-probability-of-treatment-weighting estimator for $E(Y_{dt})$. After similar manipulation, $E[Y_t(\tilde{\omega}_{1t} - \tilde{\omega}_{0t})]$ becomes

$$E\left(\frac{Y_t 1(D_t = 1)}{P(D_t = 1)} - \frac{Y_t 1(D_t = 0) P(D_t = 1|J, X_t)}{P(D_t = 0|J, X_t) P(D_t = 1)}\right),$$

which is the inverse-probability-of-treatment-weighting form of the ATT.

4.4 Estimation

The identification results presented above suggest plug-in strategies for implementing latent-class matching and reweighting estimators of average causal effects. These estimators follow naturally by replacing population expressions with sample analogs formed by substituting consistent estimates for population parameters and sample averages for population moments. I present exact expressions for the estimators in Appendix B.

Latent-class matching estimators of the ATE and ATT can be obtained as follows. First, either modal- or proportional-assignment-based estimates of covariate×latent-class-specific average counterfactual outcomes can be constructed from sample analogs of the expressions for $E_{Y_{dt}|J,x_t}$ given in Propositions 1 and 2. Second, differences between treated and untreated counterfactual outcomes can be used to estimate (X_t, J)-specific average causal effects, which can be aggregated to the population and treated-population level by replacing population moments with sample analogs in expressions (15) and (16).

Latent-class reweighting estimators of the ATE and ATT based on either modal or proportional assignment can be implemented similarly. The weights ω_{dt} and $\tilde{\omega}_{dt}$ can be estimated using sample analogs of the expressions for them given in Proposition 3. The ATE and ATT can then be estimated as sample averages of observed outcomes multiplied by these estimated weights.

The matching and reweighting estimators differ only in how they aggregate from the individual to the population and treated-population levels. An advantage of the matching estimators is that they produce initial estimates of covariate×latent-class-specific average counterfactual outcomes. These may be of intrinsic interest, particularly when the latent classes can be interpreted meaningfully.[13] An advantage of the reweighting estimators is that they may be combined with regression and other methods in order to obtain doubly robust estimators as in Robins et al. (1994).

4.4.1 Overlap

The latent-class matching and reweighting estimators can be implemented using either the modal or proportional assignment procedures described above. One drawback to modal assignment is that it is possible that, for some combinations of covariates and treatment status, individuals are modally assigned to one or more latent classes with probability zero. In this case, the expressions used to recover covariate×latent-class-specific average counterfactual outcomes under modal assignment (Proposition 1) are not well defined, while the corresponding expressions for proportional assignment (Proposition 2) are. However, this problem may indicate poor overlap between the characteristics of treated and untreated individuals, which can complicate estimation under proportional assignment as well.

The overlap component (2) of Assumption 1, which holds that individuals from all covariate×latent-class strata are treated with probability strictly between zero and one, ensures that average counterfactual outcomes are identified within those strata. In estimation, poor overlap (i.e., treatment probabilities close to zero or one) may lead to imprecision due to numerical difficulties associated with computing either the uncorrected modal- or proportional-assignment estimators (i.e., the sample analogs of expressions (8) and (12)) and inverting the corresponding misclassification probability matrices. As a rule of thumb for addressing poor overlap in selection-on-observables research designs, Crump et al. (2009) propose estimating average treatment effects among those for whom the probability of receiving the treatment conditional on covariates is between .1 and .9. This approach can be adapted to the latent-class setting by excluding, when aggregating to the population or

[13] For example, Haviland and Nagin (2005) assign individuals to three latent trajectory groups characterized by chronic, low, and declining levels of delinquency.

treated-population level, (X_t, J) strata for which estimates of $P(D_t = 1|x_t, j)$ lie outside of the interval $(.1, .9)$.

4.4.2 Inference

The consistency of the latent-class matching and reweighting plug-in estimators follows under Assumptions 1 and 2 and standard regularity conditions (see Newey and McFadden, 1994, Theorem 2.6) from the consistency of the MLE (or of GMM) and Slutsky's theorem on probability limits of functions of random variables (Wooldridge, 2010, Lemma 3.4).[14] The simplest way to conduct statistical inference using the estimators, which involve somewhat complex functions of the estimated parameters of the auxiliary model, is with a nonparametric bootstrap in which both the auxiliary model and average causal effects are estimated at each replication.

In addition, the reweighting estimators can be viewed as method-of-moments estimators for which large-sample results are readily available. Let β denote either the ATE or ATT, θ denote the parameters of the likelihood function for the finite-mixture model, and $\omega_t = \omega_t(\theta)$ denote $\omega_{1t} - \omega_{0t}$ or $\tilde{\omega}_{1t} - \tilde{\omega}_{0t}$ (depending on which average effect is being estimated), themselves functions of θ. Then putting

$$m(\beta, \theta, Z_i, W_i, Y_{it}) = \left(\beta - Y_{it}\omega_{it}(\theta), \frac{\partial \log[\ell(Z_i, W_i|W_{1i}, \theta)]}{\partial \theta} \right),$$

the reweighting estimators satisfy $N^{-1} \sum_{i=1}^{N} m(\hat{\beta}, \hat{\theta}, Z_i, W_i, Y_{it}) = 0$. Putting $G = E(mm')$ and $H = E[\partial m / \partial(\beta, \theta)]$, Proposition 4 follows from Theorem 6.1 of Newey and McFadden (1994).[15]

Proposition 4. *Let β denote the time-t ATE or ATT, and $\hat{\beta}$ a corresponding latent-class reweighting estimate. Under Assumptions 1 and 2,*

$$\sqrt{N}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, v),$$

where v is the first element of $H^{-1}GH^{-1'}$.

[14] Consistency also requires that the auxiliary model is correctly specified and identified.

[15] Although the modal assignment estimators are not differentiable because of their dependence on the $1(\hat{J} = j)$, the result only requires differentiability in a neighborhood of the true θ_0 (Newey and McFadden, 1994, Theorem 3.4), which will hold if the $\hat{J}$ are unique so that the probability limits of the estimators are well defined. The proportional assignment estimators are differentiable everywhere.

5 Monte Carlo study

5.1 Data-generating processes

I conduct two simulation studies to illustrate the use of the latent-class matching estimators and to provide evidence on their small-sample performance. In the settings that I simulate, the treatment is available in each time period and, though successive treatment decisions are serially correlated, those decisions are independent of one another conditional on covariates and latent class membership.

For both studies, the J are drawn from a three-point discrete distribution with mass function given by the vector $p_j = (.3, .5, .2)$ for $j \in \{1, 2, 3\}$. For the first study, the X are time-invariant draws from a four-point discrete distribution with conditional probability mass functions

$$p_{x|j} = \begin{cases} (.25, .25, .25, .25) & \text{if } j = 1 \\ (.2, .2, .3, .3) & \text{if } j = 2 \\ (.1, .2, .3, .4) & \text{if } j = 3 \end{cases},$$

for $x \in \{1, 2, 3, 4\}$. For the second, I assume that the X are time invariant and distributed independently of J with mass function $p_x = (.2, .3, .3, .2)$. For both studies, counterfactual outcomes are determined by

$$Y_{0t} = J + X + \epsilon_{0t} \quad \text{and} \quad Y_{1t} = 1 + 2J + 2X + \epsilon_{1t}$$

and treatment status is determined according to

$$D_t = 1(1 + 2J + X + \epsilon_t > 5),$$

where $\epsilon_{dt} \sim N(0, 2)$ and $\epsilon_t \sim N(0, 4)$ for $d_t \in \{0, 1\}$ and $t \in \{1, \ldots, T\}$, so that observed outcomes are given by $Y_t = D_t Y_{1t} + (1 - D_t) Y_{0t}$.

I selected these specifications to ensure overlap across (X_t, J) strata and to generate differences in strata-specific causal effects. While neither time effects nor lagged variables affect treatment status or counterfactual outcomes in the models that I use to generate the data, the methods developed above could be applied in such circumstances.

5.2 Estimation

For each study, I simulate 250 datasets $\{D_{it}, Y_{1it}, Y_{0it}, Y_{it}, X_i, J_i\}$ for $i \in \{1, \ldots, 2000\}$ and $t \in \{1, \ldots, 10\}$. I then compute latent-class matching estimates of the population, treated

population, and latent-class-specific average effects of the treatment. I implement these estimators from the perspective of an empiricist who hypothesizes, correctly, that the data are generated from the processes described above, but only has access to the variables $\{D_{it}, Y_{it}, X_i\}$ that would be observed in applications.

The repeated-treatment setting suggests implementing the first step of the procedure using a finite-mixture model of the sequence of treatment decisions in which successive decisions are independent of one another conditional on covariates and latent-class membership.[16] In the notation of Section 3, I estimate a model of the form in (5), with $Z = (D_1, \ldots, D_T)$ and $W = X$. This approach illustrates that the first step of the estimation procedure can be implemented using a relatively simple finite-mixture model.

For each simulated dataset, I estimate a finite-mixture logit with individual conditional likelihood function

$$\begin{aligned} \ell(d_1, \ldots, d_T | x; \gamma, \beta) &= \sum_{j=1}^{|J|} P(J = j|x) \prod_{t=1}^{T} P(D_t = d_t | x, j) \\ &= \sum_{j=1}^{|J|} \left(\frac{e^{\gamma_{xj}}}{\sum_{k=1}^{|J|} e^{\gamma_{xk}}} \right) \prod_{t=1}^{T} \left(\frac{e^{d_t \beta_{xj}}}{1 + e^{\beta_{xj}}} \right), \end{aligned} \tag{17}$$

where the γ_{x1} are normalized to zero. Estimates of the auxiliary model based on (17), which allows separate coefficients for every combination of x and j, are completely nonparametric. I estimate the parameters of the model, and from them the posterior probabilities q_j of latent-class membership, using the interface to the EM algorithm provided by the R package FlexMix (Grün and Leisch, 2008a). In testing on a preliminary simulated dataset, using three latent classes maximized both the BIC and the AIC, and I estimate the models under this constraint.

I then use the estimated parameters of the auxiliary model to compute the latent-class matching estimators described in Section 4. I construct the estimators using proportional assignment to latent classes in order to allow for the possibility that overlap is poor in some of the simulates. Because outcomes and the treatment decision are stationary over time in the environment that I simulate, I estimate causal effects that are averaged over time (in addition to covariates and latent-class membership) by pooling observations from all time periods when computing the matching estimators (rather than estimating separate treatment effects for each time period).

For the purpose of comparison, I also estimate average causal effects using three other

[16] There are multiple ways that finite-mixture models of the observed variables could be used as part of the first step of the procedure. For example, the first step might also consist of treatment-status-specific finite-mixture models of discretized measures of outcomes observed in the first few periods, as in (6).

methods. The first of these, feasible only in a simulation setting, is a J-observed exact matching estimator that computes (X, J)-specific average treatment effects using the sample analogs of $E(Y|X, J, D = 1) - E(Y|X, J, D = 0)$, then aggregates those estimated effects to the population, treated population, and latent-class levels. The resulting estimates can be considered the truth against which the other methods are measured. I also present results for an uncorrected proportional-assignment matching estimator that computes (X, J)-specific average causal effects as sample analogs of (12), then aggregates to higher levels. Finally, I use exact observed-covariate matching, which aggregates X-specific causal effects computed as the sample analogs of $E(Y|X, D = 1) - E(Y|X, D = 0)$ to the population and treated population levels, ignoring J entirely.

5.3 Results

Figure 1 presents a graphical summary of the distributions of the estimates for the first simulation study, in which X and J are correlated (the means and standard deviations of the estimates are tabulated in Appendix D).[17] The latent-class matching estimates of the ATE (labelled "LC") are centered around the median (and as Table 1 shows, mean) estimate obtained by exact matching, treating J as an observed covariate (labelled "Obs"). Moreover, the interquartile range of the latent-class matching estimates is only slightly larger than when J is observed, suggesting minimal loss of precision. In contrast, the uncorrected-latent-class and observable-covariate matching estimates (labelled "Unc." and "Cov.," respectively), while less dispersed, are centered around medians that overstate the effect of the treatment.

The estimates of the ATT show a similar pattern. The distributions of the J-observed and latent-class matching estimates have similar medians, though the latent-class matching estimator is less precise. Uncorrected latent-class matching again overstates the average effect. In this case, observed-covariate matching approximates the ATT well, though below I provide evidence that this is because, for this data-generating process, the correlation between X and J makes the covariates good proxies for covariate×latent-class strata among the treated population.

The figure also shows the distributions of estimated latent-class-specific average treatment effects. Here, the latent-class matching estimates are much less precise in comparison to both latent-class estimates of the (overall) ATE and ATT and to J-observed and uncorrected latent-class matching estimates of the latent-class-specific treatment effects.[18] Intuitively

[17] In these plots, the "whiskers" indicate the minimum and maximum values (with outliers excluded according to R's default algorithm), the boxes indicate the interquartile range, and the solid vertical lines indicate the median.

[18] A limitation of finite-mixture models is that they are only identified up to arbitrary permutations of the labels $j \in \{1, \ldots, |J|\}$. To reduce the loss of precision due to inter-simulation label swapping in the J-specific

this is because, though covariate×latent-class-specific counterfactual outcomes are not well-identified when data are sparse in a given covariate stratum, aggregation according to the empirical covariate distribution mitigates the resulting uncertainty.[19] Despite this, the latent-class estimates are all centered near the median J-observed estimate. In contrast, while the uncorrected latent-class matching estimates are more precise, their interquartile range never contains the J-observed median.

The last panel of Figure 1 summarizes estimates of the unconditional latent-class distribution (the vertical dotted lines show the assumed population proportions). The auxiliary model identifies this distribution well, some mean reversion in the estimated proportions notwithstanding. Though this is not surprising given the identification results discussed in Section 3.1, it is worth emphasizing that this distribution is estimated entirely from repeated observations of treatment status within covariate strata, highlighting the power of panel variation in observables to identify models with unobserved heterogeneity.

Figure 2 presents the results of the second simulation study, which differs from the first only in that J is distributed independently of X.[20] The relative performance of the estimators is much the same as in the previous study: the latent-class matching estimates are closely centered around the median J-observed matching estimates. Though the latent-class matching estimates are somewhat less precise than the uncorrected-latent-class and observed-covariate matching estimates, the latter tend to overstate the average casual effects in question. Unlike in the first study, the observed-covariate matching estimates of the ATT are not centered around the J-observed median estimate, presumably because in this case observed covariates are less informative about latent-class membership conditional on treatment status. As before, the finite-mixture model estimates the latent-class distribution well.

estimates, I set the initial weights used by FlexMix to the observed J. This is similar to the common practice of using the full sample estimate to initialize each bootstrap estimate (which is not possible in a simulation setting since the data are redrawn before each estimation). This initialization procedure does not drive the results however, as I obtain nearly identical results using random initializations.

[19] Parenthetically, much of the dispersion between the minimal and maximal latent-class-specific average treatment effect estimates appears to be the result of a few simulates in which overlap is poor for some covariate×latent-class combinations, making some of the $\hat{Q}_{d,x}$ difficult to invert. In applications, the precision of the latent-class-specific estimates could be improved by excluding covariate×latent-class strata with poor overlap.

[20] I obtain similar results when I repeat the first simulation study with a misspecified likelihood function that assumes that J is distributed independently of X and uses smooth linear functional forms for the logit models (in which case the misspecification is the assumption of logistic errors).

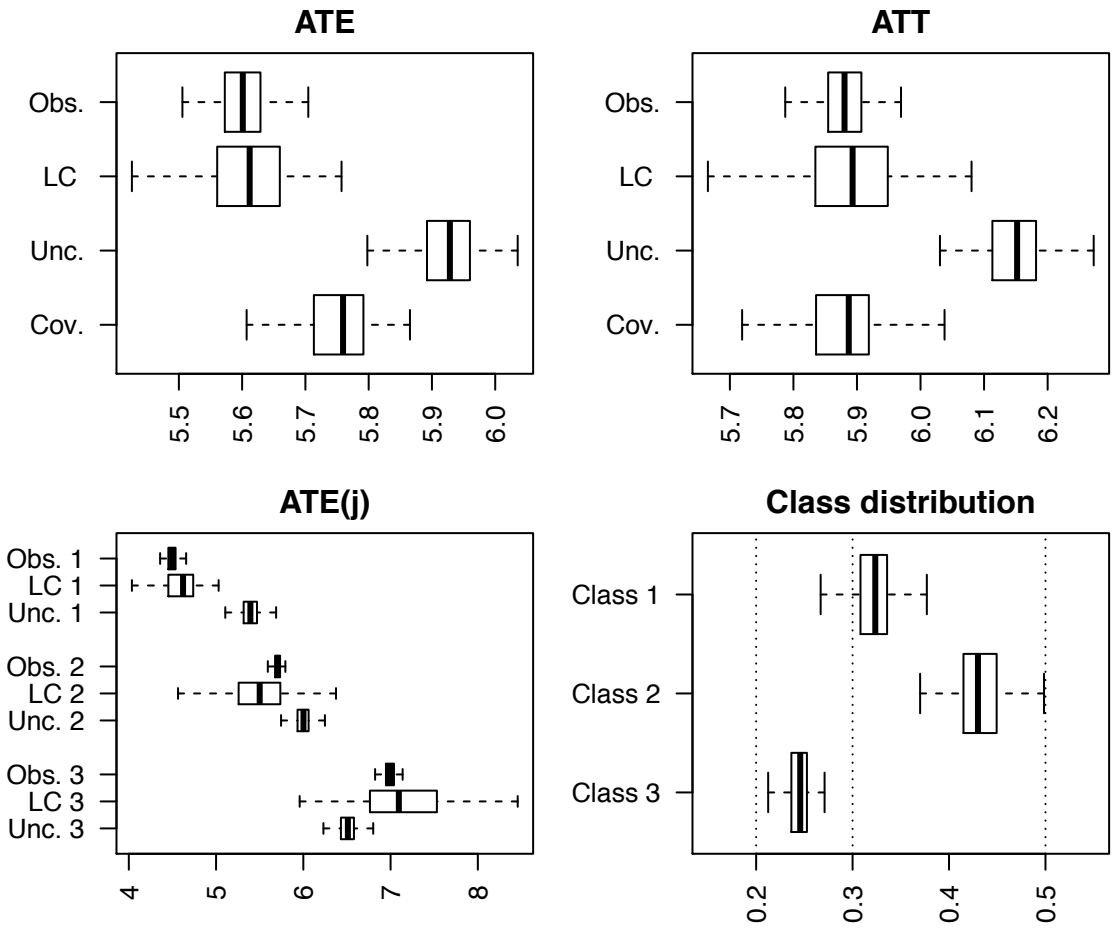

Figure 1: Estimates for simulation study 1. "Obs." denotes matching on covariates where J is observed, "LC" denotes the latent-class matching esitmator, "Unc." denotes the uncorrected latent-class matching estimator, and "Cov." denotes matching on observed covariates when J is unobserved.

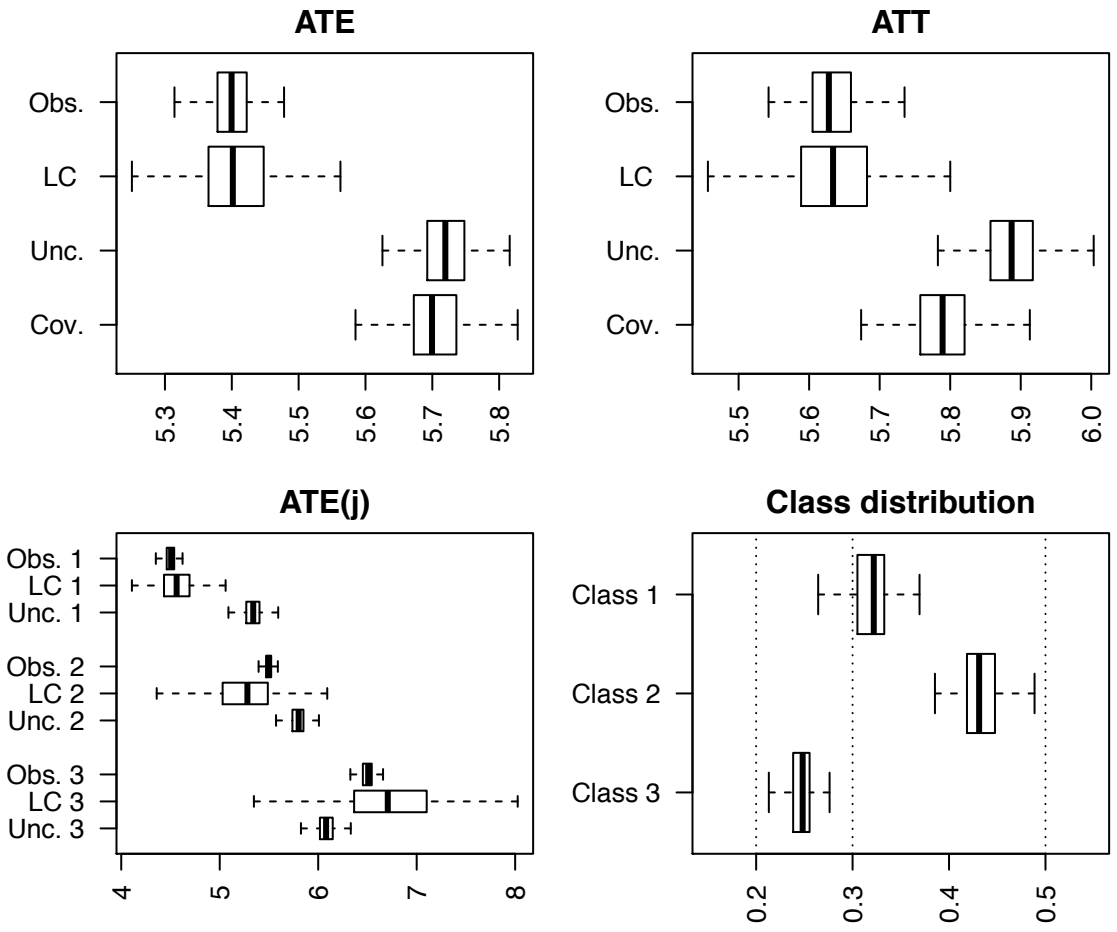

Figure 2: Estimates for simulation study 2. "Obs." denotes matching on covariates where J is observed, "LC" denotes the latent-class matching esitmator, "Unc." denotes the uncorrected latent-class matching estimator, and "Cov." denotes matching on observed covariates when J is unobserved.

6 Conclusion

The methods developed in this paper extend nonparametric matching and propensity-score reweighting methods to settings where counterfactual outcomes are independent of treatment status conditional on observed covariates and membership in latent classes. Unlike difference-in-differences matching and reweighting, these methods place no restrictions on the relationship between counterfactual outcomes and the unobserved variables that influence treatment status. They do, however, require specification and identification of a finite-mixture model that relates a set of observed auxiliary variables to latent-class membership. The latent-class matching and reweighting estimators motivated by this approach, which circumvent estimating the distributions of counterfactual outcomes, are computationally attractive and perform well in Monte Carlo studies.

The use of finite-mixture models of panel variation in observed variables to identify and estimate causal effects in the presence of unobserved heterogeneity shows promise; this literature calls for advancement. While the methods developed in this paper can accommodate continuous variables in the auxiliary and causal models through semiparameric discretization, future work should adapt or develop non- and semiparametric methods to reduce the complexity of modeling continuous variables. More generally, new identification results for mixture models, and refined ways of estimating them, will expand the circumstances under which we can use those models to conduct causal inference in the presence of unobserved heterogeneity.

A Directly modeling counterfactual outcomes

As noted above, differences in average outcomes between treated and untreated individuals assigned to the same latent classes do not identify average causal effects because, regardless of the assignment procedure used, the classes to which individuals are assigned may not be those to which they actually belong. While the identification results developed above identify average causal effects by correcting for these misclassifications, another solution to the problem is to include outcomes Y_t in the auxiliary model. The reason that directly modeling outcomes along with the other variables in the finite-mixture model solves the misclassification problem is that, under Assumptions 1 and 2, $P(Y_t|j, d_t, x_t) = P(Y_{dt}|j, x_t)$. Thus, the distributions of counterfactual outcomes are identified as part of the auxiliary finite-mixture model. These distributions, in turn, can be aggregated to identify covariate$\times$latent-class specific, population, and treated-population average causal effects.

This approach can also be used to identify average causal effects using latent-class match-

ing methods that do not need to be corrected for misclassification. When Y_t is an element of (Z, W), uncorrected proportional assignment correctly identifies average outcomes with covariate×latent-class strata. To see this, note that

$$\begin{aligned} E\left(\frac{Y_t q_j}{E(q_j|x_t,d_t)}\middle|x_t,d_t\right) &= E\left(\frac{Y_{dt}E[1(J=j)|Z,W]}{E\left\{E[1(J=j)|Z,W]\middle|x_t,d_t\right\}}\middle|x_t,d_t\right) \\ &= E\left[E\left(\frac{Y_{dt}1(J=j)}{P(J=j|x_t,d_t)}\middle|Z,W\right)\middle|x_t,d_t\right] \\ &= E(Y_{dt}|J=j,x_t), \end{aligned}$$

where the first equality follows from the law of iterated expectations and the definition of q_j, the second because (Y_t, X_t, D_t) is an element of (Z, W), and the third from Assumption 1.[21]

The direct approach can also be combined with propensity-score reweighting procedures. For example, when Y_t is included in (Z, W) and Assumptions 1 and 2 hold,

$$E\left(\frac{Y_t 1(D_t=d_t)q_j}{P(D_t=d_t|X_j,j)P(J=j)}\right) = E(Y_{dt}|J=j),$$

which can be used to identify the average effect of the treatment among those in class j (and aggregated according to the population level using the distribution of J).[22] Similarly,

$$\begin{aligned} E\left[q_j\left(\frac{Y_t D_t}{P(D_t=1|J=j)P(J=j)} - \frac{Y_t(1-D_t)P(D_t=1|X_t,j)}{P(D_t=0|X_t,j)P(D_t=1|j)P(J=j)}\right)\right] \\ = E(Y_{1t}-Y_{0t}|D_t=1,J=j), \end{aligned}$$

which can be used to identify the average effect of the treatment on treated members of class j.[23]

[21] Another way to see that proportional assignment does not identify $E(Y_{dt}|x_t, j)$ when Y_t is excluded from the auxiliary model is to note that the second equality in the above does not hold in that case.

[22] For brevity, drop the time subscripts and let $\pi_j = P(J=j)$, $P_j = P(D_t=1|X_t, J=j)$ and $1_j = 1(J=j)$. Since $q_j = E(1_j|Z,W)$ and Y is an element of (Z,W), and since $YD = Y_1 D$, we have that $E(YDq_j/P_j\pi) = E[E(Y_1 D 1_j|Z,W)/(P_j\pi_j)]$. Since P_j is a function of X (which is an element of (Z,W)), this can also be written $E\{E[Y_1 D 1_j/(P_j\pi_j)|Z,W]\} = E[Y_1 D 1_j/(P_j\pi_j)]$, or using the law of total probability, $E(Y_1 D/P_j|j)$. Iterating expectations and using Assumption 1, this can be expressed $E[E(Y_1 D/P_j|X,j)|j] = E[E(Y_1|X,j)E(D|X,j)/P_j|j] = E(Y_1|j)$, where the last equality follows from the definition of P_j. The case for Y_0 follows by replacing D with $(1-D)$.

[23] Extending the notation and arguments in the previous footnote, let $P_{1j} = P(D=1|j)$ and note first that $E(YDq_j/P_1\pi_1) = E\{E[Y_1 D 1_j/(P_{1j}\pi)|Z,W]\} = E(Y_1|D=1,j)$.

Next, write $E\{Y(1-D)q_j P_j/[(1-P_j)P_{1j}\pi_j]\} = E\{E[Y_1(1-D)1_j P_j|Z,W]/[(1-P_j)P_{1j}\pi_j]\} = E\{E[Y_1(1-D)1_j P_j/[(1-P_j)P_{1j}\pi_j]|Z,W\} = E\{Y_1(1-D)1_j P_j/[(1-P_j)P_{1j}\pi_j]\}$. Applying the law of total probability, this last expression becomes $E\{Y_1(1-D)P_j/[(1-P_j)P_{1j}]|j\}$. Iterating expectations and using the inde-

B Expressions for latent-class matching and reweighting estimators

The estimated parameters $\hat{\theta}$ of the auxiliary finite-mixture model can be used to estimate the posterior probabilities q_{ji} that observations $i \in \{1, \ldots, N\}$ belong to latent classes $j \in \{1, \ldots, |J|\}$ as

$$\hat{q}_{ji}(\hat{\theta}) = \frac{\ell(Z_i, W_i | J = j, W_{1i}, \hat{\theta})}{\sum_{k=1}^{|J|} \ell(Z_i, W_i | J = k, W_{1i}, \hat{\theta})},$$

and the modal latent-class assignments $\hat{J}_i$ as $\hat{J}_i(\hat{\theta}) = \text{argmax}_{i \in \{1,\ldots,|J|\}} \hat{q}_{ji}(\hat{\theta})$.

B.1 Latent-class matching estimators

Let $\hat{a}_{ji}(\hat{\theta}) \in \{1[\hat{J}_i(\hat{\theta}) = j], \hat{q}_{ji}(\hat{\theta})\}$ denote an estimated modal or proportional assignment of observation i into latent class j. Following Propositions 1 and 2, the (j, k)th element of the associated matrix $A_{d_t,x_t} \in \{P_{d_t,x_t}, Q_{d_t,x_t}\}$ of misclassification probabilities can be estimated as

$$\hat{A}_{x_t,d_t}(j, k) = \frac{\sum_{\{i:X_{it}=x_t, D_{it}=d_t\}} \hat{q}_{ki}(\hat{\theta}) \hat{a}_{ji}(\hat{\theta})}{\sum_{\{i:X_{it}=x_t, D_{it}=d_t\}} \hat{q}_{ji}(\hat{\theta})}$$

for all $(x_t, d_t) \in \text{supp}(X_t, D_t)$. The (uncorrected for misclassification) modal- and proportional-assignment approaches impute counterfactual outcomes using estimates of $E^A_{Y_t|d_t,x_t} \in \{E_{Y_t|\hat{J},d_t,x_t}, E_{Y_t q_J|d_t,x_t}\}$, formed by stacking

$$\hat{E}^A_{Y_t|d_t,x_t}(j) = \frac{\sum_{\{i:X_{it}=x_t, D_{it}=d_t\}} Y_{it} \hat{a}_{ji}(\hat{\theta})}{\sum_{\{i:X_{it}=x_t, D_{it}=d_t\}} \hat{a}_{ji}(\hat{\theta})}$$

for each $j \in \{1, \ldots, |J|\}$. Following Propositions 1 and 2, the vectors $E_{Y_{dt}|J,d_t,x_t}$, $d_t \in \{0, 1\}$, of (X_t, J)-specific average counterfactual outcomes can be estimated under either modal or proportional assignment using

$$\hat{E}_{Y_{dt}|J,d_t,x_t} = \hat{A}^{-1}_{d_t,x_t} \hat{E}^A_{Y_t|d_t,x_t}$$

in order to estimate covariate×latent-class-specific average treatment effects $\hat{ATE}_t(x_t, j) = \hat{E}_{Y_{1t}|J,1,x_t} - \hat{E}_{Y_{1t}|J,0,x_t}$. These can then be aggregated to the population and treated-population

pendence of Y_1 and D given X and J twice, this can be written $E\{E[Y_1(1-D)P_j/[(1-P_j)P_{1j}]|X, j]|j\} = E\{P_j E[Y_1|X, j]E[(1-D)|X, j]P_j/[(1-P_j)P_{1j}]|j\} = E[E(DY_1|X, j)/P_{1j}|j] = E(Y_1|D = 1, j)$.

levels following (15) and (16) as

$$\hat{ATE}_t = \sum_{x_t,j} \left(\frac{\sum_{\{i:X_{it}=x_t\}} \hat{q}_{ji}(\hat{\theta})}{N} \right) \hat{ATE}_t(x_t, j)$$

and

$$\hat{ATT}_t = \sum_{x_t,j} \left(\frac{\sum_{\{i:X_{it}=x_t, D_{it}=1\}} \hat{q}_{ji}(\hat{\theta})}{\sum_{i=1}^{N} D_{it}} \right) \hat{ATE}_t(x_t, j).$$

B.2 Latent-class reweighting estimators

Following Proposition 3, the weights ω_{dit}, $d_t \in \{0, 1\}$, can be estimated as

$$\hat{\omega}_{dit}(\hat{\theta}) = \sum_j \sum_k \hat{A}^{-1}_{d_t, X_{it}}(j,k) \frac{\hat{a}_{ki} 1(D_{it} = d_t) \left(\frac{\sum_{\{i':X_{i't}=X_{it}\}} \hat{q}_{ji'}(\hat{\theta})}{\sum_{\{i':X_{i't}=X_{it}\}} 1} \right)}{\left(\frac{\sum_{\{i':X_{i't}=X_{it}, D_{i't}=D_{it}\}} \hat{a}_{ki'}(\hat{\theta})}{\sum_{\{i':X_{i't}=X_{it}, D_{i't}=D_{it}\}} 1} \right) \left(\frac{\sum_{\{i':X_{i't}=X_{it}\}} D_{i't}}{\sum_{\{i':X_{i't}=X_{it}\}} 1} \right)}.$$

Note that the set $\{i' : X_{i't} = X_{it}\}$ refers to individuals $i' \in \{1, \ldots, N\}$ with the same realization of X as individual i in period t. Similarly, $\tilde{w}_{1it}$ can be estimated as

$$\hat{\tilde{\omega}}_{1it} = \frac{D_{it}}{\left(\frac{\sum_{\{i':X_{i't}=X_{it}\}} D_{it}}{\sum_{\{i':X_{i't}=X_{it}\}} 1} \right)}$$

and $\tilde{w}_{0it}$ as

$$\hat{\tilde{\omega}}_{0it}(\hat{\theta}) = \sum_j \sum_k \hat{A}^{-1}_{1, X_{it}}(j,k) \frac{Y_{it} \hat{a}_{ki} (1 - D_{it}) \left(\frac{\sum_{\{i':X_{i't}=X_{it}, D_{i't}=1\}} \hat{q}_{ji'}(\hat{\theta})}{\sum_{\{i':X_{i't}=X_{it}, D_{i't}=1\}}} \right)}{\left(\frac{\sum_{\{i':X_{i't}=X_{it}, D_{i't}=0\}} \hat{a}_{ki'}(\hat{\theta})}{\sum_{\{i':X_{i't}=X_{it}, D_{i't}=0\}} 1} \right) \left(\frac{\sum_{\{i':X_{i't}=X_{it}\}} (1 - D_{i't})}{\sum_{\{i':X_{i't}=X_{it}\}} 1} \right)} \frac{\left(\frac{\sum_{\{i':X_{i't}=X_{it}\}} D_{i't}}{\sum_{\{i':X_{i't}=X_{it}\}} 1} \right)}{\left(\frac{\sum_{i'=1}^{N} D_{i't}}{N} \right)}.$$

The time-t ATE can then be estimated as $\hat{ATE}_t = N^{-1} \sum_{i=1}^{N} (\hat{\omega}_{1it} - \hat{\omega}_{0it}) Y_{it}$ and the ATT estimated as $\hat{ATT}_t = N^{-1} \sum_{i=1}^{N} (\hat{\tilde{\omega}}_{1it} - \hat{\tilde{\omega}}_{0it}) Y_{it}$.

C Proof of Proposition 3

Proof. To prove the first part, note that time-t mean counterfactual outcomes satisfy

$$E(Y_{dt}) = E\left(\sum_j E(Y_{dt} | X_t, j) P(j | X_t) \right)$$

$$
\begin{aligned}
&= E\left[\sum_j \sum_k A^{-1}_{d_t,X_t}(j,k) E\left(\frac{Y_t a_k}{E(a_k|X_t,d_t)}\middle| X_t, d_t\right) P(j|X_t)\right] \\
&= E\left[E\left(\sum_j \sum_k A^{-1}_{d_t,X_t}(j,k)\frac{Y_t a_k}{E(a_k|X_t,d_t)} P(j|X_t)\middle| X_t, d_t\right)\right] \\
&= E\left[E\left(\sum_j \sum_k A^{-1}_{d_t,X_t}(j,k)\frac{Y_t a_k 1(D_t=d_t)}{E(a_k|X_t,d_t)}\frac{P(j|X_t)}{P(d_t|X_t)}\middle| X_t\right)\right] \\
&= E(Y_t \omega_{dt}),
\end{aligned}
$$

where the second equality follows from Propositions 1 and 2, the third follows because A_{d_t,X_t} and $P(j|X_t)$ are functions of X_t, the fourth from the law of total probability, and the fifth from the law of iterated expectations.

For the second part, clearly $E[Y_t 1(D_t = 1)/P(D_t = 1)] = E(Y_{1t}|D_t = 1)$. Under Assumptions 1 and 2,

$$
\begin{aligned}
E(Y_{0t}|D_t=1) &= \sum_{x_t}\sum_j E(Y_{0t}|x_t,j)P(j|x_t,D_t=1)P(x_t|D_t=1) \\
&= \sum_{x_t}\sum_j E(Y_{0t}|x_t,j)P(j|x_t,D_t=1)\frac{P(x_t|D_t=1)}{P(x_t)}P(x_t) \\
&= E\left(\sum_j E(Y_{0t}|X_t,j)P(j|X_t,D_t=1)\frac{P(D_t=1|X_t)}{P(D_t=1)}\right) \\
&= E\left(\sum_j\sum_k A^{-1}_{1,X_t}(j,k)\frac{E(Y_t a_k|D_t=0,X_t)}{E(a_k|D_t=0,X_t)}P(j|X_t,D_t=1)\frac{P(D_t=1|X_t)}{P(D_t=1)}\right) \\
&= E\left[E\left(\sum_j\sum_k A^{-1}_{1,X_t}(j,k)\frac{Y_t a_k P(j|X_t,D_t=1)}{E(a_k|D_t=0,X_t)}\frac{P(D_t=1|X_t)}{P(D_t=1)}\middle| X_t, D_t=0\right)\right] \\
&= E\left[E\left(\sum_j\sum_k A^{-1}_{1,X_t}(j,k)\frac{Y_t a_k 1(D_t=0)P(j|X_t,D_t=1)}{E(a_k|D_t=0,X_t)P(D_t=0|X_t)}\frac{P(D_t=1|X_t)}{P(D_t=1)}\middle| X_t\right)\right] \\
&= E(Y_t\tilde{\omega}_{0t}),
\end{aligned}
$$

where $x_t \in \operatorname{supp} X_t$ and $j,k \in \{1,\ldots,|J|\}$. In the above, the first three equalities follow from basic probability calculus, the fourth from Propositions 1 and 2, the fifth because A_{d_t,X_t} and $P(D_t = 1|X_t)$ are functions of X_t, and the six and seventh from the laws of total probability and, respectively, iterated expectations. □

D Simulation study summary tables

Table 1: Summary of estimates for simulation study 1

(a) Treatment effects

	ATE	ATT	ATE(J=1)	ATE(J=2)	ATE(J=3)
J observed	5.60	5.88	4.50	5.71	6.99
	(0.04)	(0.04)	(0.06)	(0.04)	(0.06)
Latent-class matching	5.61	5.88	4.57	5.39	7.32
	(0.08)	(0.10)	(0.28)	(0.84)	(1.77)
Uncorrected latent-class matching	5.92	6.15	5.38	5.99	6.52
	(0.05)	(0.05)	(0.14)	(0.11)	(0.13)
Observed covariate matching	5.75	5.88			
	(0.05)	(0.05)			

(b) Latent-class distribution

	P(J=1)	P(J=2)	P(J=3)
True	0.3	0.5	0.2
Estimated	0.32	0.44	0.24
	(0.04)	(0.05)	(0.02)

Notes—Means and standard deviations from 250 simulations. Elements of the latent-class distribution estimated as the unconditional means of the estimated priors. Other entries are described in the main text.

Table 2: Summary of estimates for simulation study 2

(a) Treatment effects

	ATE	ATT	ATE(J=1)	ATE(J=2)	ATE(J=3)
J observed	5.40	5.63	4.50	5.50	6.50
	(0.03)	(0.04)	(0.06)	(0.04)	(0.07)
Latent-class matching	5.40	5.63	4.53	4.92	7.42
	(0.07)	(0.09)	(0.29)	(2.30)	(4.50)
Uncorrected latent-class matching	5.72	5.89	5.33	5.79	6.09
	(0.05)	(0.05)	(0.12)	(0.10)	(0.12)
Observed covariate matching	5.70	5.79			
	(0.05)	(0.05)			

(b) Latent-class distribution

	P(J=1)	P(J=2)	P(J=3)
True	0.3	0.5	0.2
Estimated	0.31	0.44	0.24
	(0.03)	(0.04)	(0.02)

Notes—Means and standard deviations from 250 simulations. Elements of the latent-class distribution estimated as the unconditional means of the estimated priors. Other entries are described in the main text.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*, 1–19.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*(1), 113–132.

Aguirregabiria, V. and P. Mira (2016). Identification of games of incomplete information with multiple equilibria and unobserved heterogeneity. Working paper.

Allman, E. S., C. Matias, and J. A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics 37*(6A), 3099–3132.

Arcidiacono, P. and J. B. Jones (2003). Finite mixture distributions, sequential likelihood and the EM algorithm. *Econometrica 71*(3), 933–946.

Arcidiacono, P. and R. A. Miller (2011). CCP estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica 79*(6), 1823–1867.

Bartolucci, F., L. Grilli, and L. Pieroni (2012a). Estimating dynamic causal effects with unobserved confounders: A latent class version of the inverse probability weighted estimator. Working paper.

Bartolucci, F., L. Grilli, and L. Pieroni (2012b). Inverse probability weighting to estimate causal effects of sequential treatments: A latent class extension to deal with unobserved confounding. Working paper.

Bolck, A., M. Croon, and J. Hagenaars (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis 12*(1), 3–27.

Bonhomme, S., K. Jochman, and J.-M. Robin (2016). Nonparametric estimation of finite mixtures. *Journal of the Royal Statistical Society, Series B 76*(1), 211–229.

Bray, B. C., S. T. Lanza, and X. Tan (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling 22*(1), 1–11.

Compiani, G. and Y. Kitamura (2016). Using mixtures in econometric models: A brief review and some new results. *Econometrics Journal 19*, C95–127.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

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

Grün, B. and F. Leisch (2008a). Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software 8*(4).

Grün, B. and F. Leisch (2008b). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification 25*(2), 225–247.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics 31*(1), 201–224.

Haviland, A., D. S. Nagin, P. R. Rosenbaum, and R. E. Tremblay (2008). Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Developmental Psychology 44*(2), 422–436.

Haviland, A. M. and D. S. Nagin (2005). Causal inferences with group based trajectory models. *Psychometrika 70*(3), 557–578.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–654.

Heckman, J. J. and B. Singer (1984a). The identifiability of the proportional hazard model. *Review of Economic Studies 51*(2), 231–241.

Heckman, J. J. and B. Singer (1984b). A method of minimizing the distributional impact in econometric models for duration data. *Econometrica 52*(2), 271–320.

Henry, M., Y. Kitamura, and B. Salanié (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics 5*(1), 123–144.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Horvitz, D. G. and D. G. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Hu, Y. and M. Shum (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics 171*, 32–44.

Imbens, G. (2014). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Kasahara, H. and K. Shimotsu (2009). Nonparametric identification and estimation of finite mixture models of dynamic discrete choices. *Econometrica 77*(1), 135–175.

Kasahara, H. and K. Shimotsu (2014). Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society 76*(1), 97–111.

Lanza, S. T., D. L. Coffman, and S. Xu (2013). Causal inference in latent class analysis. *Structural equation modelling 20*(3), 361–383.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 36, pp. 2112–2245. Elsevier Science.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics 34*(4), 1265–1269.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis 18*(4), 450–469.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*(1), 95–103.