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

John Gardner*

Abstract

This paper extends 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. I apply the methods to estimate the effect of gang membership on violent delinquency.

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

JEL Codes: C21, C22.

*Department of Economics, University of Mississippi. Email: jrgardne@olemiss.edu. I thank John Conlon, Natalia Kolesnikova and Alice Sheehan, the editor, associate editor, and three anonymous referees for helpful comments and suggestions.

1 Introduction

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 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 therefore do not represent the causal effect of the treatment on outcomes.

Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (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. In addition, Bonhomme and Sauder (2011) show that the difference-in-differences approach can be generalized to recover the entire distribution of counterfactual outcomes.

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

From these assumptions, identification of average causal effects proceeds in two steps. First, 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 probabilities of membership in each latent class. The purpose of this part of the procedure is to impute the latent classes to which individuals belong, according to their posterior probabilities of class membership (given their realizations of the observed auxiliary variables). While the identification procedure does not require that the data have a panel dimension, the evolution of observed variables over time provides a natural source of variables to be included in the auxiliary model. 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.

Second, observed-covariate$\times$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.[1] These average counterfactual outcomes can then be used to identify covariate$\times$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). This approach to identification suggests non- and semi-parametric latent-class matching and reweighting estimators for average causal effects under selection on unobservables. Because the estimators obviate the need to estimate the full distribution of counterfactual outcomes, they are computationally attractive. They are also simple enough to be implemented using routines available in standard statistical packages.

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 how latent variables affect counterfactual outcomes. It can be applied when the effect of those variables on outcomes changes over time (which would invalidate the parallel-trends assumption) or when counterfactual outcomes are not additively separable in latent class membership (as required by the method of Bonhomme

[1]The insight that the classification probabilities can be recovered from elements of the finite-mixture model is originally due to work by Bolck et al. (2004) on latent-class regression (also see Vermunt, 2010).

and Sauder 2011). 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 those variables reveal information about the 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 they 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, Nagin, Rosenbaum, and Tremblay (2008), and Bartolucci, Grilli, and Pieroni (2012a,b).[2] 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×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

[2] A related literature analyzes the sensitivity of treatment effect estimates to the assumption of selection on observables by estimating those effects under different assumptions about an unobserved covariate (Rosenbaum and Rubin, 1983a; Imbens, 2003). The method developed here differs from these papers by using information from the auxiliary model to relate individuals' latent classes to their observables in order to point-identify average causal effects.

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.

I introduce the identification and estimation procedures by example in Section 2. In Section 3.1, I discuss the assumptions and data requirements of the procedures. In Section 3.2, I give an overview of the specification, identification and estimation of finite-mixture models. I show how latent-class-specific average counterfactual outcomes are identified in Section 3.3 and how they can be used to identify average causal effects in Section 3.4. In Section 4, I provide Monte-Carlo evidence on the small-sample performance of the estimators. In Section 5, I apply the results to estimate the effect of gang membership on violent delinquency when unobserved delinquent tendencies may influence the decision to join a gang. I conclude in Section 6.

2 An introductory example

To motivate the identification procedures developed in this paper, I begin with a simple example based on work by Haviland and Nagin (2005) and Haviland et al. (2008), who study the effect of gang membership on violently delinquent behavior under a selection-on-observables assumption which holds that potentially counterfactual delinquent behavior with and without gang membership is independent of observed gang membership conditional on histories of such behavior. Their key innovation is to identify the effect of gang membership by comparing behavioral outcomes among gang members and non-members assigned to the same latent trajectory groups, a set of unobserved classes that characterize individuals' propensity for delinquency. They impute the groups to which individuals belong using a finite-mixture model which assumes that successive realizations of delinquent behavior during a pre-treatment period are independent of one another conditional on group membership, and show that gang members and non-members assigned to the same latent group have similar behavioral histories.

A natural extension of this idea is to identify the effect of gang membership on delinquency when individuals select into gang membership on the basis of their *unobserved* trajectory-group membership. Let Y_{0T} represent the delinquent behavior that an individual would exhibit at time T if they were not a gang member in that period, Y_{1T} represent their behavior if they joined a gang in that period, and $D_T \in \{0, 1\}$ be an indicator for gang membership in period T, so that observed behavior at time T can be expressed as $Y_T = (1 - D_T)Y_{0T} + D_T Y_T$. Suppose for simplicity that each individual belongs to one of two latent classes $J \in \{1, 2\}$, taken for the purpose of this application to be behavioral trajectory groups, and

that counterfactual outcomes (Y_{1T}, Y_{0T}) are independent of gang membership D_T conditional on trajectory-group membership J.

In addition, suppose that we also observe histories of delinquent behavior $Z = (Y_1, \ldots, Y_{T-1})$ recorded during a pre-treatment period in which nobody has joined a gang and, following Haviland and Nagin (2005) and Haviland et al. (2008), that successive realizations of such behavior are independent of one another conditional on trajectory-group membership. I assume for simplicity that these measures are discrete (though the argument that follows applies without change to discretized versions of continuous variables, and can be modified to accommodate the continuous case).[3] In this case, regardless of eventual gang membership, the likelihood of an individual's history for these auxiliary variables can be expressed as

$$\ell(Z|D_T = d) = \sum_{j \in \{1,2\}} P(J = j|D_T = d) \prod_{t=1}^{T-1} P(Y_t = y_t|J = j, D_T = d), \tag{1}$$

for $d \in \{0, 1\}$.

Because identification and estimation of finite-mixture models of the form in (1) is well understood (I discuss these issues in Section 3.2), suppose that the components $P(J = j|D_T = d)$ and $P(Y_t = y_t|J = j, D_T = d)$ for $j \in \{1, 2\}$, $d \in \{0, 1\}$, and $y_t \in \operatorname{supp} Y_t$, of (1) are known. Using these components, the posterior probability q_j that an individual belongs to trajectory group $j \in \{1, 2\}$ is identified from Bayes' rule as

$$q_j = P(J = j|Z, D_T) = \frac{P(J = j|D_T) \prod_{t=1}^{T-1} P(Y_t|J = j, D_T)}{\sum_{k \in \{1,2\}} P(J = k|D_T = d) \prod_{t=1}^{T-1} P(Y_t|J = k, D_T)}.$$

Given these posteriors, a straightforward way to impute the trajectory groups to which individuals belong is to assign each individual to the group $\hat{J}$ for which their posterior is largest. Under the assumptions on the data-generating process, an intuitive approach to identifying the causal effect of gang membership among members of group j is to compare time-T outcomes between gang members and non-members assigned to group j, that is via the comparison

$$E(Y_T|D_T = 1, \hat{J} = j) - E(Y_T|D_T = 0, \hat{J} = j). \tag{2}$$

Similar approaches have been taken elsewhere in the literature (see, for example, Haviland and Nagin, 2005; Haviland et al., 2008; Bartolucci et al., 2012a,b).[4]

[3] As I discuss in Section 3.2, the identification argument can also be modified to allow for more complex relationships among the auxiliary variables.

[4] It is important to note that Haviland and Nagin (2005) and Haviland et al. (2008) use this approach under the assumption of selection on observables and show that treated and untreated units assigned to the same class have similar observables.

The problem with this approach in the current context is that the latent trajectory groups to which individuals are assigned are not necessarily those to which they actually belong; imputed group membership is an error-ridden measure of true membership. Furthermore, because this misclassification problem is one of prediction, rather than sampling, error, it will persist even in large samples or at the population level. A consequence of this is that mean outcomes among those with the same gang-membership status and imputed trajectory group do not identify trajectory-group-specific counterfactual outcomes. In fact, under the ancillary assumption that counterfactual outcomes (Y_{0T}, Y_{1T}) are independent of the behavioral histories Z conditional on trajectory-group membership J (in other words, the behavioral histories are only informative about counterfactual outcomes insofar as they are correlated with trajectory-group membership), we can write

$$\begin{aligned} E(Y_T|D_T = d, \hat{J} = j) &= E(Y_T|D_T = d, J = 1, \hat{J} = j)P(J = 1|\hat{J} = j, D_T = d) \\ &\quad + E(Y_T|D_T = d, J = 2, \hat{J} = j)P(J = 2|\hat{J} = j, D_T = d) \\ &= E(Y_{dT}|J = 1)P(J = 1|\hat{J} = j, D_T = d) \\ &\quad + E(Y_{dT}|J = 2)P(J = 2|\hat{J} = j, D_T = d), \end{aligned} \tag{3}$$

where the second equality follows from the assumptions that counterfactual outcomes are independent of treatment status and behavioral histories given group membership. Thus, gang-membership×imputed-trajectory-group-specific mean outcomes identify a weighted average of counterfactual outcomes for members of all trajectory groups.

In addition to showing that (2) does not identify the average effect of the treatment among members of trajectory group j, (3) suggests a strategy for recovering that effect from observable variables and the components of the finite-mixture model. The classification probabilities $P(J = k|\hat{J} = j, D_T = d)$, $j, k \in \{1, 2\}$, in (3) are identified via

$$\begin{aligned} P(J = k|\hat{J} = j, D_T = d) &= \frac{\sum_{z \in \operatorname{supp} Z} P(J = k|\hat{J} = j, z, d)P(\hat{J} = j|z, d)P(z|d)}{P(\hat{J} = j|d)} \\ &= \frac{\sum_z P(J = k|z, d)1(\hat{J} = j)P(z|d)}{P(\hat{J} = j|d)} = E\left(\frac{q_k 1(\hat{J} = j)}{P(\hat{J} = j|d)}\middle| d\right), \end{aligned}$$

where the first equality follows from Bayes' rule, the second because $\hat{J}$ is a function of the components of the finite-mixture model, and hence of (Z, D), and the third from the definition of the posteriors. Consequently, the system formed by stacking (3) for $j \in \{1, 2\}$ can be solved to identify the mean trajectory-group-specific counterfactual outcomes $E(Y_{dT}|J = j)$ for each $d \in \{0, 1\}$ and $j \in \{1, 2\}$, and hence the trajectory-group-specific causal effects

$E(Y_{1T} - Y_{0T}|J = j)$ of gang membership on violent delinquency.[5]

Together, these arguments suggest a plug-in matching estimator for latent-class-specific average causal effects, formed by replacing the population objects in equations (1) and (3) with sample analogs from the data and an estimated finite-mixture model. These class-specific causal effects can then be aggregated to the population and treated-population level to estimate the population average treatment effect and average effect of the treatment on the treated. The resulting estimator is straightforward to implement, requires little more than estimation of a relatively simple finite-mixture model (many popular statistical packages, including R, Stata, and SAS include facilities for estimating such models) and solving a small system of equations, and obviates the need for computationally intensive estimation of the full distribution of latent-class-specific counterfactual outcomes.[6] The remainder of this paper generalizes this simple example to more complex settings and presents evidence on the performance of the estimators.

3 Identification and estimation

3.1 Preliminaries

Generalizing the example of Section 2, let D_{it} be an indicator for whether individual i receives a (binary) treatment at time t and Y_{dit} denote the possibly counterfactual outcome that i would experience if assigned to treatment status $d_t \in \{0, 1\}$ 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

[5] In the present case with two latent classes, Cramer's rule gives

$$E(Y_{dT}|J = j) = \frac{E(Y_T|\hat{J} = j, d)P(J = k|\hat{J} = k, d) - E(Y_T|\hat{J} = k, d)P(J = k|\hat{J} = j, d)}{P(J = 1|\hat{J} = 1, d)P(J = 2|\hat{J} = 2, d) - P(J = 1|\hat{J} = 2, d)P(J = 2|\hat{J} = 1, d)},$$

for $j, k \in \{1, 2\}$ and $j \neq k$.

[6] An alternative approach is to estimate the latent-class-specific distributions of counterfactual outcomes Y_{dT}, and hence causal effects, by including observed outcomes Y_T in the finite mixture model. While this approach is conceptually simpler, it poses several challenges. It may be unclear whether the finite-mixture model is identified when it is augmented to include observed outcomes. Modeling observed outcomes also increases the dimension, and hence the computational complexity, of the identification and estimation problems, particularly when observed outcomes are continuous or high-dimensionally discrete. Finally, including a continuous outcome of interest in the finite-mixture model either requires discretization, choosing a parametric distribution, or using relatively sophisticated smoothing methods (all of which may introduce approximation error or require making somewhat arbitrary modeling decisions).

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 (I assume for simplicity that the observed covariates X_{it} are either discrete or discretized versions of continuous variables, although the arguments below can be modified to accommodate continuous covariates). Following Rosenbaum and Rubin (1983b), strong ignorability in this context means that counterfactual outcomes are independent of treatment status conditional on observed covariates and latent class membership (this is also known as conditional independence), and that the characteristics of treated and untreated individuals overlap in the sense that the probability of receiving the treatment conditional on covariates and latent class membership is strictly between zero and one (this is also known as overlap).[7] Formally:

Assumption 1. *Counterfactual outcomes and treatment assignment satisfy*

$$(Y_{0it}, Y_{1it}) \perp\!\!\!\perp D_{it} | X_{it}, J_i, \tag{4}$$

and

$$P(D_{it} = 1 | X_{it}, J_i) \in (0, 1), \tag{5}$$

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

If latent class membership were observed, Assumption 1 would be identical to the requirements for causal identification using standard matching and inverse probability weighting estimators that assume selection on observables. Although, in contrast to those settings, the latent-class overlap component (5) of Assumption 1 is not directly verifiable, as I discuss in Section 3.5, it can be assessed as part of the identification procedure developed below. Assumption 1 is also similar to the requirements 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. Although the assumption that the unobserved determinants of treatment status and counterfactual outcomes are drawn from a discrete distribution is restrictive, as I discuss in Section 3.2, causal effect estimates obtained using the method developed below

[7] This is similar to the notion of latent ignorability in Frangakis and Rubin (1999), in which counterfactual outcomes are independent of randomized treatment assignment conditional on unobserved treatment-compliance groups (complier, always taker, never taker, and defier).

can be interpreted as approximations that improve with the number and dimension of the auxiliary variables modeled in the first step of the procedure.

The identification procedure requires that the data contain observations on a set of auxiliary dependent variables $Z_i = (Z_{i1}, \ldots, Z_{iS})$ and possibly a set of auxiliary covariates $W_i = (W_{i1}, \ldots, W_{iS})$, 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 (I assume these variables are also discrete or discretized, though this is not essential to the argument). The purpose of these variables is to identify the components of a finite-mixture model in order to impute the latent classes to which individuals belong.[8] In the example of Section 2, the auxiliary variables are the pre-treatment-period histories of delinquent behavior, and treatment status is the only auxiliary covariate, though the procedure could be repeated within covariate strata if counterfactual delinquency were only independent of gang membership conditional on trajectory-group membership and a set of observed covariates.

The identification procedure also requires that the auxiliary model meet the following conditions:

Assumption 2.

$$(Y_{0it}, Y_{1it}) \perp\!\!\!\perp Z_i | X_{it}, J_i, \tag{6}$$

and X_{it} *and* D_{it} *are elements of* (Z_i, W_i).

Unlike Assumption 1, which places restrictions on the underlying data-generating process, Assumption 2 places restrictions on the auxiliary model used to identify average counterfactual outcomes. The first part of Assumption 2 requires that the auxiliary variables Z_i exert no influence on counterfactual outcomes after conditioning on the observable covariates X_{it} from the causal model and latent class membership J_i (this is the sense in which the elements of Z_i can be considered auxiliary variables). This requirement is necessary because the identification procedure requires that average counterfactual outcomes are independent of predicted latent class membership (and hence of the auxiliary variables used to predict that membership) conditional on actual latent class membership.[9] In many studies predicated on conditional independence between counterfactual outcomes and treatment status, this assumption will be natural. The introductory example of Section 2 is motivated by an assumption that counterfactual behavioral outcomes are independent of gang membership

[8]The identification procedure can be applied with no time dimension whatsoever if there is a set of auxiliary variables and covariates that are measured contemporaneously with treatment status and outcomes. I develop the argument in terms of a panel structure because variation in observables over time suggests a natural source of auxiliary variables and covariates to use in the finite-mixture model.

[9]This is illustrated in the introductory example of Section 2 as well as the proofs of Propositions 1 and 2 below.

conditional on latent trajectory-group membership. In this setting, it is natural to suppose that counterfactual outcomes are also independent of the auxiliary pre-treatment-period behavioral histories used to assign individuals to trajectory groups. Indeed, the failure of this assumption would call into question the underlying premise that latent trajectory-group membership is the only factor that influences both counterfactual behavioral outcomes and gang membership.

However, the first part of Assumption 2 should not be assumed uncritically to follow from Assumption 1. For example, it is possible that histories of delinquent behavior influence counterfactual outcomes in a way that does not affect the decision to join a gang (in which case Assumption 1 would hold without conditioning on such histories, while Assumption 2 would not). In this case, histories of pre-treatment-period delinquency would be a poor choice of auxiliary variables to use in the first-step finite-mixture model. The plausibility of Assumption 2 ultimately depends on the setting of study and the candidate auxiliary variables. As in research designs based on selection on observables, the validity of Assumptions 1 and 2 should be evaluated carefully on the basis of theoretical reasoning and prior empirical evidence.

The second part of the assumption requires simply that treatment status D_{it} and the covariates X_{it} from the causal model are included in the auxiliary model. This requirement is necessary because the components of the auxiliary model are used to correct for errors in latent class assignments which, like average counterfactual outcomes, may depend on covariates and treatment status.[10]

While the second part of this assumption allows for the possibility that the set Z_i of auxiliary variables contains elements that do not also appear in the causal model (i.e., Y_{it}, D_{it}, and X_{it}), it does not require such elements. Nor does the assumption require that there are auxiliary covariates, although it does allow for them. For example, in the Monte Carlo exercise presented in Section 4, I use an auxiliary model of the history $Z_i = (D_{i1}, \ldots, D_{iT})$ of the treatment decision over time as a function of the same time-invariant covariate $W_i = X_i$ included in the causal model to analyze the effect of a treatment that can be received in multiple periods. On the other hand, in the empirical application of Section 5, I use an auxiliary model of behavioral histories $Z_i = (Y_{i0}, \ldots, Y_{i3})$ that are measured in a pre-treatment period before anyone has joined a gang (and are therefore not included in the causal model), with treatment status D_i as an auxiliary covariate, to estimate the effect of gang-membership on delinquent behavior in a post-treatment period. Assumption 2 accommodates both of these configurations.

[10]The proofs of Propositions 1 and 2 below make this intuition rigorous.

3.2 The auxiliary model

The first step of the identification procedure involves using an auxiliary finite-mixture model to impute the latent classes to which individuals belong. Although mixtures models have a long history in econometrics (Heckman and Singer, 1984a,b, e.g.), a growing literature on the conditions under which finite-mixture models are identified without distributional assumptions or other parametric restrictions (see Hall and Zhou, 2003; Allman, Matias, and Rhodes, 2009; Kasahara and Shimotsu, 2009; Hu and Shum, 2012; Henry, Kitamura, and Salanié, 2014; Bonhomme, Jochman, and Robin, 2016; Compiani and Kitamura, 2016) has found increasing use in econometric applications involving unobserved heterogeneity (see, for example, Arcidiacono and Jones, 2003; Arcidiacono and Miller, 2011; Aguirregabiria and Mira, 2016). This section provides a brief overview of the specification, identification and estimation of finite-mixture models in order to demonstrate the applicability of the approach developed in this paper to a variety of different settings.

Finite mixtures are not, in general, nonparametrically identified without restrictions on how the auxiliary variables and auxiliary covariates are related. However, the requirements for identification have been established for several important and flexible classes of finite mixture models. Nonparametric identification is important in this context because it implies that the components of the finite-mixture model used to assign observations to latent classes, and ultimately the causal effects of the treatment, are identified from variation in the observed auxiliary variables and covariates.[11]

3.2.1 Latent-class models

One case where identification is well understood is when the auxiliary dependent variables Z_s, $s \in \{1, \ldots, S\}$, are independent of one another conditional latent class membership J and potentially a set of time-invariant auxiliary covariates W. This is the type of model used in the introductory example of Section 2, which models realizations Z of pre-treatment period delinquent behavior as a function of treatment status D_T at time T.

When the auxiliary variables are discrete, finite mixtures of this form are known as latent-class models. Assume for now that the number $|J|$ of latent classes is known (I discuss identification and estimation of $|J|$ below). In this case, the likelihood of observing $Z = z$ given that $W = w$ can be expressed as

$$\ell(z|w) = \sum_{j=1}^{|J|} P(J = j|w) \prod_{s=1}^{S} P(Z_s = z_s|j, w). \tag{7}$$

[11] Parametric identification of finite mixtures follows more readily (see Teicher, 1963; Grün and Leisch, 2008b, for a discussion).

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.[12]

3.2.2 Dynamic discrete-choice models

In some cases, it may be unrealistic to assume that auxiliary variables are independent of one another conditional on latent class membership and the auxiliary covariates, as in the simple structure of the latent-class model (7). For example, Bartolucci et al. (2012a,b) estimate the effect of wage subsidies on employment using an inverse-probability-of-treatment-weighting estimator based on propensity scores that depend on latent class membership. To assign firms to latent classes, they use a multi-period finite-mixture model in which the receipt of wage subsidies (Z_s) in each period depends on latent-class membership as well as a vector W_s of time-varying firm characteristics (employment, wages, capital, sales, profits and prior receipt of subsidies), which in turn depends on lagged values of these characteristics.

More complex models like theirs can be viewed as dynamic discrete choice models with unobserved heterogeneity, for which a number of nonparametric identification results have been established. For example, if the auxiliary covariates evolve according to a first-order Markov process, the likelihood of observing a sequence z of auxiliary variables given a sequence w of auxiliary covariates is

$$\begin{aligned}\ell(z, w|w_1) = \sum_{j=1}^{|J|} P(J = j|w_1) \prod_{s=2}^{S} P(Z_s = z_s|j, w_s) P(W_s = w_s|j, w_{s-1}, z_{s-1}) \\ \times P(Z_1 = z_1|j, w_1).\end{aligned} \tag{8}$$

Kasahara and Shimotsu (2009, also see Hu and Shum, 2012) establish conditions under which several such models are nonparametrically identified, depending on the characteristics of the data (the number of periods and dimension of the covariates) and the structure of the model (for example, whether the distributions of the auxiliary variables are stationary, whether lagged auxiliary variables are included among the auxiliary covariates, and whether the transitions between auxiliary covariates depend on latent class membership).[13] These

[12]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). 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 (also see Kasahara and Shimotsu, 2009; Bonhomme et al., 2016, for the continuous case).

[13]In general, they show that the number of latent classes for which the model is identified depends on the dimension of the covariates, while additional time periods help identify models with features such as

identification results for dynamic finite-mixture models allow the causal effect identification procedure developed below to be implemented using first-step models in which the auxiliary variables and covariates are related through more complex structures than the simple latent-class model (7).

3.2.3 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 unrestricted components of the finite-mixture model—that is, the $P(j|w_1)$, $P(z_s|j, w_s)$ and $P(w_s|j, w_{r<s}, z_{r<s})$ for $j \in \{1, \ldots, |J|\}$, $(z_s, w_s) \in \text{supp}(Z_s, W_s)$, and $s \in \{1, \ldots, S\}$. While the discussion in this section has centered on completely nonparametric identification, these components can also be specified parametrically in order to implement the latent-class matching estimators semiparametrically (this is analogous to implementing an inverse-probability-weighting estimator using propensity scores obtained using a flexible parametric logit or probit model). 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).[14] Many statistical packages include some facility for estimating finite-mixture models.

Another consideration in estimating finite mixtures is the number of latent 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) (most routines for estimating finite mixtures include these statistics as part of their output by default; 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.

While the formal results on identification of average causal effects developed below require that the number of latent classes for which the finite-mixture model is identified coincides with the number of classes for which Assumption 1 holds, there is no guarantee that this will hold in any given application. However, as the results discussed above show, the number of classes for which auxiliary model is identified is increasing in the number and dimension

nonstationarity, latent-class-specific transitions between auxiliary covariates, and lagged dependent variables as covariates (see Kasahara and Shimotsu, 2009, for details).

[14] 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).

of the auxiliary variables. Consequently, causal effects obtained using the identification procedure developed below can be interpreted as approximations whose quality improves with the richness of the auxiliary variables.

3.3 Identifying average counterfactual outcomes within unobserved strata

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 the components of the auxiliary model via Bayes' rule as

$$q_{ji} = P(J_i = j|Z_i, W_i) = \frac{P(J_i = j|W_{1i})\ell(Z_i, W_i|j, W_{1i})}{\sum_{k=1}^{|J|} P(J_i = k|W_{1i})\ell(Z_i, W_i|k, W_{1i})}, \tag{9}$$

and are therefore identified along with the auxiliary model.[15]

Two-step analyses based on finite-mixture models (such as those by Haviland and Nagin 2005, Haviland et al. 2008, and Bartolucci et al. 2012a,b) 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. Such studies typically use one of two procedures for assigning observations to latent classes. 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.

As I note above, because the classes to which individuals are assigned under these procedures may not be the classes to which they actually belong, estimators that compare treated and untreated individuals assigned to the same latent class do not identify latent-class-specific average causal effects. However, the following results show that under either assignment method the posterior probabilities of latent class membership can be used to recover the distributions of these classification errors, and ultimately correct for them in order to identify average causal effects within unobserved strata.

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

3.3.1 Modal assignment

In a typical modal-assignment analysis, covariate×latent-class specific average counterfactual outcomes are imputed as

$$E(Y_t|\hat{J} = j, d_t, x_t). \tag{10}$$

Although (10) does not identify the mean counterfactual outcome $E(Y_{dt}|J = j, x_t)$ of interest, Proposition 1 below shows how this outcome (and hence average causal effects) can be recovered by correcting for errors in the latent-class assignments. The proof of the proposition is nearly identical to the argument developed in the example of Section 2. Under Assumptions 1 and 2, the modal assignment estimand (10) represents an average of counterfactual outcomes across all latent classes, weighted by the probabilities of membership in each class conditional on modal assignment to class j. Since these probabilities are identified from the auxiliary model, the system of equations formed by stacking these weighted averages for each possible latent-class assignment $\hat{J} \in \{1, \ldots, |J|\}$ can be solved to recover the average causal effects of interest. All proofs are presented in Appendix A.

Proposition 1 (Identification via modal assignment)**.** *Let $E_{Y_{dt}|J,x_t}$ be the $|J|$-vector of average counterfactual outcomes with jth element $E(Y_{dt}|J = j, X_t = x_t)$. Let $E_{Y_t|\hat{J},d_t,x_t}$ be the $|J|$-vector of modal assignment estimands with jth element $E(Y_t|\hat{J} = j, D_t = d_t, X_t = x_t)$. Let P_{d_t,x_t} be the $|J| \times |J|$ matrix of classification probabilities with (j,k)th element $P(J = k|\hat{J} = j, D_t = d_t, X_t = x_t)$.*

Under Assumptions 1 and 2,

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

for all (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

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

Embedded in Propositions 1 is the ancillary requirement that the matrix of classification probabilities P_{d_t,x_t} be invertible for each treatment status d_t and covariate stratum x_t. Of note, this restriction rules out the use of auxiliary variables that are independent of latent class membership, in which case the finite-mixture model is not identified (a model where all individuals belong to the same class and one where the parameters do not vary with class membership would have the same likelihood). In such cases, the classification-probability matrices are not invertible (because the latent-class assignments $1(\hat{J} = j)$ are independent of

J) and, consequently, average counterfactual outcomes are not identified.[16] This invertibility requirement may also be violated if the overlap component of Assumption 1 fails, a possibility that I discuss further in Section 3.5.

3.3.2 Proportional assignment

A similar argument can be applied to methods based on proportional assignment to latent classes. Proportional assignment methods classify individuals as members of each latent class in proportion to their posterior probabilities of class membership, and impute counterfactual outcomes within (X_t, J) strata as

$$E\left(\frac{Y_t q_j}{P(J=j|x_t,d_t)}\middle|x_t,d_t\right) = E\left(\frac{Y_t q_j}{E(q_j|x_t,d_t)}\middle|x_t,d_t\right). \tag{11}$$

Although the proportional assignment method is motivated by uncertainty about latent class membership, the resulting estimand still does not identify average counterfactual outcomes within covariate×latent-class strata. Proposition 2 shows that a similar error correction can be used to recover these mean counterfactual outcomes. Like that for the previous proposition, the proof shows that the proportional assignment estimand identifies a weighted average of latent-class-specific mean counterfactual outcomes, which can be inverted to recover the mean counterfactual outcomes of interest themselves.

Proposition 2 (Identification via proportional assignment)**.** *Let $E_{Y_{dt}|J,x_t}$ be the $|J|$-vector of average counterfactual outcomes with jth element $E(Y_{dt}|J=j,X_t=x_t)$. Let $E_{Y_t q_J|d_t,x_t}$ be the $|J|$-vector of proportional assignment estimands with jth element $E(Y_t q_j|D_t=d_t,X_t=x_t)/E(q_j|D_t=d_t,X_t=x_t)$. Let Q_{d_t,x_t} be the $|J|\times|J|$ matrix of expected proportional latent-class assignments 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}

[16] As I note in Section 3.1, Assumption 2 will be violated if some of the auxiliary variables are correlated with counterfactual outcomes conditional on latent class membership and the covariates included in the causal model (for example, if delinquent behavior at time $T-1$ were correlated with counterfactual behavior at time T). A natural solution to such violations is to include the offending auxiliary variables among the covariates X_{it} (in which case counterfactual outcomes would be conditionally independent of these variables by construction). However, including *all* of the auxiliary variables among those covariates would violate the invertibility requirements (both the modal and proportional classification matrices would consist of identical rows). It would therefore be impossible, for example, to condition on the entire history of delinquent behavior modeled in the first step of the introductory example in Section 2.

satisfies

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

The above caveat regarding the importance of the invertibility requirement on the classification-probability matrices for Proposition 1 applies mutatis mutandis to Proposition 2, with Q_{d_t,x_t} in place of P_{d_t,x_t} and $P(J=j|Z)$ in place of $1(\hat{J}=j)$.

3.4 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 $E(Y_{dt}|x_t, j)$ 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

$$ATE_t(x_t, j) = E(Y_{1t} - Y_{0t}|x_t, j). \tag{12}$$

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

$$ATE_t = E(Y_{1t} - Y_{0t}) = \sum_{x_t,j} ATE_t(x_t, j)P(X_t = x_t, J = j), \tag{13}$$

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

$$ATT_t = E(Y_{1t} - Y_{0t}|D_t = 1) = \sum_{x_t,j} ATE_t(x_t, j)P(X_t = x_t, J = j|D_t = 1), \tag{14}$$

where $(x_t, j) \in \text{supp}(X_t, J)$.[17] The aggregation weights used in (13) and (14) are identified from the auxiliary model and observed covariates and treatment status by iterating expectations on $q_j = P(J = j|Z, W)$ as $P(x_t, j) = E(q_j|x_t)P(x_t)$ and $P(x_t, j|D_t = 1) = E(q_j|x_t, D_t = 1)P(x_t|D_t = 1)$.

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

[17] 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.

the characteristics of the treated and untreated populations (see, e.g., Robins, Rotnitzky, and Zhao, 1994; Hirano, Imbens, and Ridder, 2003; Abadie, 2005; Imbens and Wooldridge, 2009; Imbens, 2014). The following result shows that a similar interpretation is available for the latent-class methods developed above.

Proposition 3 (Identification via reweighting)**.** *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 classification 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 $d_t \in \{0, 1\}$,

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

and

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

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.[18]

3.5 Estimation and inference

The identification results presented above suggest plug-in strategies for implementing latent-class matching and reweighting estimators for the ATE and ATT. The strategies follow

[18]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

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

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.

naturally by replacing population objects with consistent estimates. Both estimators require (i) using the estimated finite-mixture model to estimate the posterior probabilities of latent class membership and (ii) using the estimated posteriors to estimate either the modal- or proportional-assignment classification probability matrices according to the expressions given in Propositions 1 or 2.

The latent-class matching estimator can then be implemented by (iii) applying the estimated classification probability matrices to uncorrected modal- or proportional-assignment estimators to obtain estimates of covariate×latent-class-specific average counterfactual outcomes according to the expressions in Proposition 1 for modal assignment or Proposition 2 for proportional assignment, and (iv) aggregating covariate×latent-class-specific average treatment effects to the population and treated-population levels using estimates of the aggregation weights in expressions (13) and (14) to estimate the ATE and ATT.[19] The reweighting estimators can be implemented by (iii') combining the estimated posteriors and classification probability matrices with the data to estimate the weights for the ATE or ATT according to the expressions given in Proposition 3 and (iv') taking weighted averages of observed outcomes.

The matching and reweighting estimators are numerically equivalent. An advantage of the matching estimators is that they produce initial estimates of covariate×latent-class-specific average counterfactual outcomes, which may be of intrinsic interest when the latent classes can be interpreted meaningfully.[20] 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).

An important concern in any nonparametric treatment effect estimation setting is whether there is overlap between the characteristics of treated and untreated individuals. The overlap component (5) 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

[19]This amounts to matching each treated individual to every untreated individual with the same covariates and latent class membership. The simplest way to accommodate continuous covariates is via semiparametric discretization. In addition, if the components of the finite-mixture model are specified as parametric functions of continuous covariates, the proportional-assignment classification probabilities can be estimated for each individual using sample analogs of the second equality in expression (20) of the proof of Proposition 2 in Appendix A, which can in turn be used to form a semiparametric implementation of the reweighting estimators. This is equivalent to implementing an inverse-probability-of-treatment-weighting estimator using a parametric model for the propensity score.

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

estimators (i.e., the sample analogs of expressions (10) and (11)) and inverting the corresponding classification probability matrices. The proportional-assignment-based estimators may be preferable when overlap is limited, since they can be computed even when no observations are assigned to a particular covariate×treatment×latent-class stratum (which is more likely when overlap is poor). 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 treated-population level, (X_t, J) strata for which estimates of $P(D_t = 1|x_t, j)$ lie outside of the interval $(.1, .9)$.

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).[21] The simplest way to conduct statistical inference for the latent-class matching and reweighting estimators, which involve somewhat complex functions of the data and estimated finite-mixture model, is with a nonparametric bootstrap in which both the auxiliary model and average causal effects are estimated at each replication (because finite mixtures are only identified up to permutations of the labels on the latent classes, it is good practice when bootstrapping latent-class-specific estimators to initiate estimation at each bootstrap replication with estimates obtained using the full sample; see Grün and Leisch 2008a). I present asymptotic standard errors based on a method-of-moments interpretation of the reweighting estimators in Appendix B.

4 Monte Carlo studies

4.1 Data-generating process

I conduct several 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 the main study, the J are drawn from a three-point discrete distribution with mass function given by the vector $p_j = (.3, .5, .2)$ for $j \in \{1, 2, 3\}$ and the X are time-invariant

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

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 $x \in \{1, 2, 3, 4\}$. 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 $Y_t = (1 - D_t)Y_{0t} + D_t Y_{1t}$.

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.

4.2 Estimation

For the main 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.[22] In the notation of Section 3.2, I estimate a model of the form in (7), with $Z = (D_1, \ldots, D_T)$ and $W = X$ (i.e., the auxiliary variables are the histories of the treatment decision and the auxiliary covariates are the same as the covariates in the causal model). This approach illus-

[22] 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 outcomes observed in the first few periods, as in the example of Section 2.

trates 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

$$\begin{aligned}\ell(d_1,\ldots,d_T|x;\gamma,\beta) &= \sum_{j=1}^{|J|} P(J=j|x)\prod_{t=1}^{T} P(D_t=d_t|x,j) \\ &= \sum_{j=1}^{|J|}\left(\frac{e^{\gamma_{xj}}}{\sum_{k=1}^{|J|}e^{\gamma_{xk}}}\right)\prod_{t=1}^{T}\left(\frac{e^{d_t\beta_{xj}}}{1+e^{\beta_{xj}}}\right),\end{aligned} \tag{15}$$

where the γ_{x1} are normalized to zero. Estimates of the auxiliary model based on (15), which allows separate coefficients for every combination of x and j, are completely nonparametric. I estimate the parameters of the model 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 3.5. 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 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 using sample analogs of (11), 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.

4.3 Results

Figure 1 presents a graphical summary of the distributions of the estimates for the main simulation study, in which X and J are correlated (the means and standard deviations of the estimates are tabulated in Appendix C).[23] 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 observed-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.[24]

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.[25] Intuitively 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.[26] 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.

[23] 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.

[24] In this case, observed-covariate matching approximates the ATT well, presumably because the correlation between X and J makes the covariates good proxies for covariate×latent-class strata among the treated population. In an additional study where X and J are independent, observed-covariate matching estimates the ATT poorly.

[25] To reduce the loss of precision due to inter-simulation label swapping in the J-specific 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.

[26] 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.

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 3.2, it is worth emphasizing that this distribution is estimated entirely from repeated observations of treatment status within covariate strata, highlighting the power of variation in observables to identify models with unobserved heterogeneity.

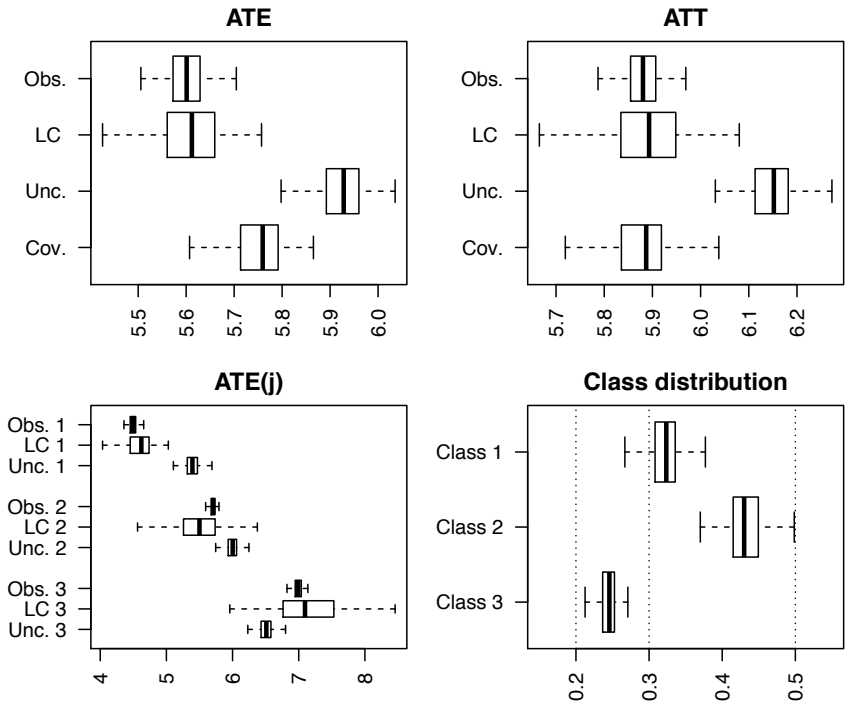

Figure 1: Estimates for main simulation study. "Obs." denotes matching on covariates where J is observed, "LC" denotes the latent-class matching esitmator, "Unc." denotes the uncorrected latent-class matching estimator, and "Cov." denotes matching on observed covariates when J is unobserved.

To provide further evidence on the performance of the estimators, I also implement several variations on this simulation study, the results of which are summarized in Table 2 of Appendix C. In the second study, the X are drawn independently from the latent classes, while the finite-mixture model still allows for the possibility that they are correlated. In the third study, the data-generating process is the same as in the main study, but the finite-mixture model is misspecified first by assuming that the J are independent of the X, and second by using a linear function form for the logits in (15) rather than a separate coefficient for each value of X (here the misspecification is really that the errors in the selection equation

are logistically distributed, since the data-generating process actually is linear). In the fourth study, the data-generating process is the same as for the main study, but the finite-mixture model is misspecified to use two latent classes instead of three. As Table 2 shows, the estimators perform well in these settings. In each case, the latent-class matching estimates are closely centered around the mean J-observed estimate, while the uncorrected-latent-class and observed-covariate matching estimators are clearly biased.[27]

5 Application

To illustrate the use of the estimators, I apply them to a setting similar to that studied in Haviland and Nagin (2005) and Haviland et al. (2008). I assume that counterfactual behavior with and without joining a gang are independent of observed gang membership conditional on latent trajectory-group membership. I then implement the procedure developed in Section 2 to estimate the causal effect of gang membership on violent delinquency under this assumption.

The data for the application are drawn from the Pathways to Desistance study (Mulvey, 2016), which follows a number of juvenile offenders, initially between the ages of 14 and 18, after conviction for a serious offense. In addition to an initial interview at the time of conviction, follow-up interviews were administered at 10 post-conviction periods (6, 12, 18, 24, 30, 36, 48, 60, 72, and 84 months past the baseline).

To analyze the effect of gang membership on violent behavior, I examine the subset of individuals who were not gang members at the time of their conviction, and had not joined a gang as of the 18-month follow-up interview. I then construct an indicator D for joining a gang between 24 and 60 months after conviction in order to examine the effect of gang membership on the probability of engaging in violently delinquent behavior between the 60- and 84-month interviews. I construct indicators for committing robbery with a weapon, shooting someone, beating someone to serious injury, being in a fight, and carrying a gun during each period. I then construct indices $Y_t \in \{0, 1, 2\}$ of violent delinquency for each pre-treatment period (the baseline and first three follow-up interviews) that take the value 0 if none of these acts were committed during period t, 1 if one such act was committed, and 2 if two or more were committed. I measure violent delinquency during the treatment

[27] To assess the asymptotic standard errors presented in Appendix B, I also conduct a study on the reweighting estimators and their standard errors. The data-generating process is the same as for the main study but to simplify the study, I only use 100 simulated datasets and only estimate the time-one ATE (for which the true value is simple to evaluate). The median estimated ATE is 5.61 and the median estimated standard error is .51. The true ATE of 5.6 was contained in all of the 95% confidence intervals, although the estimator could not be evaluated in two simulations because one of the estimated classification probability matrices was computationally singular (presumably owing to the smaller sample size).

period as a vector Y_T of indicators for having shot, robbed, beaten, or fought someone between the 60- and 84-month interviews. After dropping observations with missing values for these variables, those who were gang members prior to the 24-month interview, and those who did not complete any interviews between 24 and 60 months after the baseline, 1,140 observations remain, among which the treatment group consists of 53 individuals who joined a gang between 24 and 60 months after the baseline interview.

To assign individuals to trajectory groups, I estimate a finite-mixture model of behavioral histories $Z = (Y_0, \ldots, Y_3)$ during the pre-treatment period (that is, the initial and first three follow-up interviews) using the multinomial logit specification

$$\ell(Z|D;\theta) = \sum_{j=1}^{|J|} \pi_j \prod_{t=0}^{3} \sum_{y=0}^{2} \frac{1(Y_t = y) e^{\theta_{0yj} + \theta_{1yj} D}}{1 + \sum_{y=1}^{2} e^{\theta_{0yj} + \theta_{1yj} D}},$$

where the π_j, $j \in \{1, \ldots, |J|\}$, are the components of the latent trajectory-group distribution and the θ_{yj} are the parameters of the latent-class-specific delinquency probabilities. In the notation of Section 3.1, the auxiliary variables are $Z = (Y_0, Y_1, Y_i)$, treatment status D is included as an auxiliary covariate (W), the outcome variables are the elements of Y_T, and there are no covariates in the causal model. To determine the number of groups, I estimate the model for $|J| \in \{2, \ldots, 5\}$. The AIC and BIC are maximized at $|J| = 3$ (with five groups, the EM algorithm did not converge after 1,000 iterations). I use modal assignment to impute individuals' trajectory-group membership. Figure 2 summarizes pre-treatment-period histories of delinquent behavior for members of the three groups. The group-specific trends are similar to those reported in Haviland and Nagin (2005) using entirely different data; they label group 1 as "chronic" offenders, group 2 as "declining" offenders and group 3 as "low" offenders.[28]

The estimated propensity scores of .34, .35, and .32 for members of groups one, two, and three imply that the latent-class overlap requirement is satisfied, and hence that average causal effects are identified within latent trajectory groups. To recover these effects, I use the estimated posterior probabilities $\hat{q}_{ji}$, $j \in \{1, \ldots, |J|\}$, $i \in \{1, \ldots, N\}$, of group membership to estimate the matrices P_d, $d \in \{0, 1\}$, of modal assignment classification probabilities using the sample analogs of the expression given in Proposition 1. Following Proposition 1, I form the vector of (uncorrected) modal assignment estimates $\hat{E}_{Y_T|\hat{J},d}$ to recover estimates of the group-specific average counterfactual outcomes as $\hat{E}_{Y_{dT}|J,d} = \hat{P}_d^{-1} \hat{E}_{Y_T|\hat{J},d}$ for $d \in \{0, 1\}$. Differencing these for treated and untreated individuals produces estimates $\hat{ATE}(j)$ of the group-specific

[28] All of the group-specific trends in my data feature an initial dip not present in Haviland and Nagin (2005); this is presumably because individuals enter my dataset after an encounter with the criminal justice system.

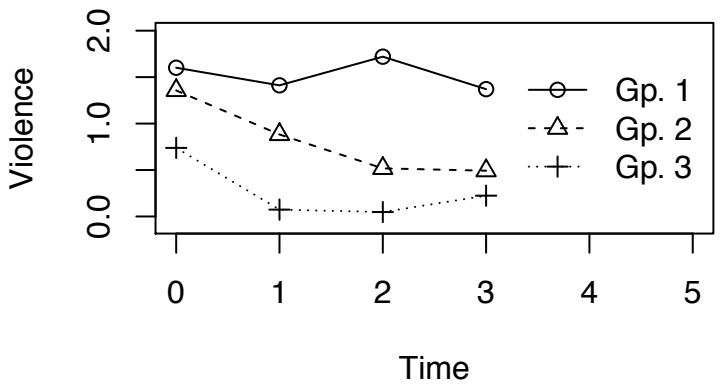

Figure 2: Group-specific behavioral histories

average causal effects, which I aggregate to the population and treated-population level as $\hat{ATE} = \sum_j \hat{ATE}(j)\hat{P}(j)$ and $\hat{ATT} = \sum_j \hat{ATE}(j)\hat{P}(j|D=1)$, where $\hat{P}(j) = N^{-1}\sum_{i=1}^{N}\hat{q}_{ji}$ and $\hat{P}(j|D=1) = \sum_{i=1}^{N}\hat{q}_{ji}D_i/(\sum_{i=1}^{N}D_i)$. For the purpose of comparison, I also compute uncorrected latent class matching estimators and naive comparisons of outcomes between gang members and non-members.

Figure 3 provides a graphical summary of the estimated ATTs of gang membership for shooting, robbing, beating, and fighting (Table 3 in Appendix C provides a detailed summary of the estimates). With the exception of shooting, the naive estimates exceed the uncorrected trajectory-group matching estimates, which themselves exceed the corrected matching estimates. Although these differences are not statistically significant (presumably because of the modest sizes of the sample and treatment group), this pattern is consistent with the notion that those with a greater inherent propensity for delinquency self-select into gang membership, and that the corrected matching estimator does a better job of controlling for this than the uncorrected estimator.

Figure 4 summarizes trajectory-group-specific average causal effects estimated by uncorrected and corrected latent trajectory-group matching. Focusing on the fighting outcome (for which the estimated ATT is significantly different from zero), the figure suggests that the uncorrected modal assignment estimator may misstate the group-specific causal effects considerably. For example, the group-2 corrected matching point estimate lies outside the 95% confidence interval for the uncorrected estimate.

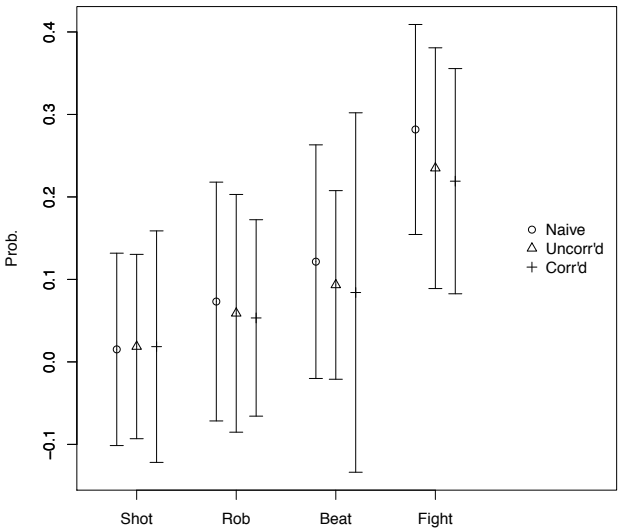

Figure 3: Estimated ATTs of gang-membership on violent delinquency. Bars show 95% confidence intervals.

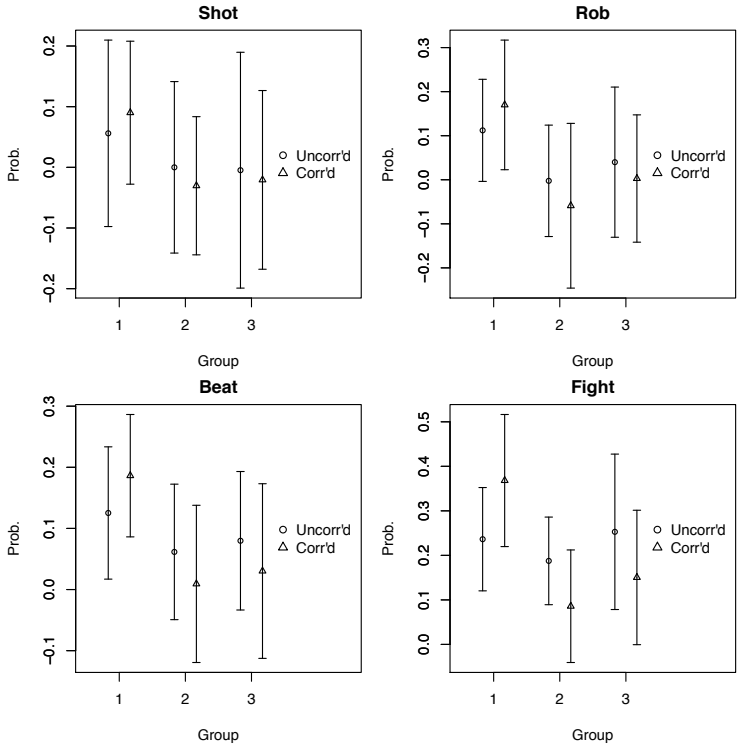

Figure 4: Estimated trajectory-group-specific effects of gang-membership on violent delinquency. Bars show 95% confidence intervals.

6 Conclusion

The methods developed in this paper extend non- and semi-parametric 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 to identify and estimate causal effects in the presence of unobserved heterogeneity shows promise; this literature calls for advancement. 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.

References

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

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.

Bonhomme, S. and U. Sauder (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics 93*, 479–494.

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.

Frangakis, C. E. and D. B. Rubin (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncomplicance and subsequent missing outcomes. *Biometrika 86*, 365–379.

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. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review Papers and Proceedings 93*(2), 126–132.

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.

Mulvey, E. P. (2016). Research on pathways to desistance [Maricopa county, AZ and Philadelphia county, PA]: Subject measures, 2000-2010]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-03-14. https://doi.org/10.3886/ICPSR29961.v2.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters 14*, 201–206.

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 (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological) 45*(2), 212–218.

Rosenbaum, P. R. and D. B. Rubin (1983b). 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.

A Proofs

Proof of Proposition 1. Modal assignment methods impute counterfactual outcomes using (10), which can be expressed via the law of total expectation as

$$E(Y_t|D_t = d_t, X_t = x_t, \hat{J} = j) = \sum_{k=1}^{|J|} E(Y_t|d_t, x_t, \hat{J} = j, J = k)P(J = k|d_t, x_t, \hat{J} = j). \quad (16)$$

Conditional on $D_t = d_t$, observed outcomes Y_t can be replaced with counterfactual outcomes Y_{dt} in the right-hand side of (16). 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:

$$E(Y_t|D_t = d_t, X_t = x_t, \hat{J} = j) = \sum_{k=1}^{|J|} E(Y_{dt}|x_t, J = k)P(J = k|d_t, x_t, \hat{J} = j). \quad (17)$$

Stacking (17) 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

$$\begin{aligned} P(J = k|x_t, d_t, \hat{J} = j) &= \frac{P(J = k, \hat{J} = j|d_t, x_t)}{P(\hat{J} = j|d_t, x_t)} \\ &= \frac{\sum_{z,w} P(J = k|\hat{J} = j, z, w)P(\hat{J} = j|z, w)P(z, w|d_t, x_t)}{P(\hat{J} = j|d_t, x_t)} \\ &= \frac{\sum_{z,w} P(J = k|z, w)1(\hat{J} = j)P(z, w|d_t, x_t)}{P(\hat{J} = j|d_t, x_t)} \\ &= E\left(\frac{q_k 1(\hat{J} = j)}{P(\hat{J} = j|d_t, x_t)}\middle| d_t, x_t\right), \end{aligned} \quad (18)$$

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 (18) 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. □

Proof of Proposition 2. Under Assumptions 1 and 2, (11) can be written

$$\begin{aligned} E\left(\frac{Y_t q_j}{P(J=j|d_t,x_t)}\middle| d_t, x_t\right) &= \sum_{z,w} E\left(\frac{Y_t q_j}{P(J=j|d_t,x_t)}\middle| z, w, d_t, x_t, \right) P(z,w|d_t,x_t) \\ &= \sum_{z,w}\left(\sum_k \frac{q_j E(Y_{dt}|J=k,z,w,d_t,x_t)}{P(J=j|d_t,x_t)} P(J=k|z,w,d_t,x_t)\right) \\ &\quad \times P(z,w|d_t,x_t) \\ &= \sum_k \left(\sum_{z,w} \frac{q_j q_k P(z,w|d_t,x_t)}{P(J=j|d_t,x_t)}\right) E(Y_{dt}|J=k,x_t), \end{aligned} \tag{19}$$

where $(z, w) \in \text{supp}(Z, W)$. The second equality in (19) 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

$$\begin{aligned} E[P(J=k|Z,W)|J=j,d_t,x_t] &= \sum_{z,w} E[P(J=k|z,w)|J=j,z,w,d_t,x_t] P(z,w|J=j,d_t,x_t) \\ &= \sum_{z,w} P(J=k|z,w) \frac{P(J=j|z,w,d_t,x_t)P(z,w|d_t,x_t)}{P(J=j|d_t,x_t)} \\ &= E\left(\frac{q_j q_k}{P(J=j|d_t,x_t)}\middle| d_t, x_t\right), \end{aligned} \tag{20}$$

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, (19) and (20) 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. □

Proof of Proposition 3. To prove the first part, note that the time-t mean counterfactual

outcomes satisfy

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

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. □

B Large-sample distribution

Let θ denote the parameters of the finite-mixture model, β the average causal effect of interest (either the ATE or ATT), and $\omega(X_t, D_t; \theta_\omega, \theta)$ the corresponding weights (that is, $\omega_1 - \omega_0$ for the ATE or $\tilde{\omega}_1 - \tilde{\omega}_0$ for the ATT), where the components θ_ω of the weights satisfy the moment conditions $E[h(X_t, D_t; \theta_\omega, \theta)] = 0$. For example, using the proportional-assignment based ATE weights, $h(X_t, D_t; \theta_\omega, \theta)$ is the vector consisting of

$$\begin{aligned}
[P(j|x_t) - q_j(\theta)]1(X_t = x_t) = 0, \quad & x_t \in \operatorname{supp} X_t, j \in \{1, \ldots, |J| - 1\} \\
[P(j|x_t, d_t) - q_j(\theta)]1(X_t = x_t)1(D_t = d_t) = 0, \quad & x_t \in \operatorname{supp} X_t, d_t \in \{0, 1\}, \\
& j \in \{1, \ldots, |J| - 1\} \\
\left(Q_{d_t, x_t}(j, k) - \frac{q_j(\theta) q_k(\theta)}{P(j|d_t, x_t)}\right) 1(X_t = x_t)1(D_t = d_t) = 0, \quad & x_t \in \operatorname{supp} X_t, d_t \in \{0, 1\}, \\
& j \in \{1, \ldots, |J|\}, k \in \{1, \ldots, |J| - 1\} \\
[P(D_t = 1|x_t) - D_t]1(X_t = x_t) = 0, \quad & x_t \in \{1, \ldots, |X_t|\}.
\end{aligned}$$

Letting $m(Z, W, Y_t; \beta, \theta_\omega, \theta) = (\beta - Y_t \omega(\theta_\omega, \theta), h(X_t, D_t; \theta_\omega, \theta), \partial \log \ell(Z, W|W_1; \theta)/\partial\theta)$, the reweighting estimators solve the sample analog of $E(m) = 0$. Putting $G = E(mm')$ and $H = E[\partial m/\partial(\beta, \theta_\omega, \theta)]$, Proposition 4, which follows from Theorem 6.1 of Newey and McFadden (1994, or from Newey, 1984), gives the asymptotic distribution of the treatment effect estimators (which can be estimated by replacing H and G with their natural sample analogs).[29]

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 $H^{-1}GH^{-1'}$.

[29] 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.

C Tables

Table 1: Summary of estimates for main simulation study

(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: Summaries of additional simulations

	Study 2				
	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)			

	Study 3				
	ATE	ATT	ATE(J=1)	ATE(J=2)	ATE(J=3)
J observed	5.60	5.88	4.50	5.70	7.00
	(0.04)	(0.04)	(0.06)	(0.04)	(0.06)
Latent-class matching	5.62	5.90	4.78	5.53	6.83
	(0.06)	(0.07)	(0.19)	(0.29)	(0.50)
Uncorrected latent-class matching	5.93	6.15	5.56	5.99	6.29
	(0.05)	(0.05)	(0.06)	(0.05)	(0.07)
Observed covariate matching	5.76	5.88			
	(0.05)	(0.05)			

	Study 4	
	ATE	ATT
J observed	5.59	5.88
	(0.03)	(0.03)
Latent-class matching	5.61	5.89
	(0.05)	(0.06)
Uncorrected latent-class matching	5.92	6.14
	(0.04)	(0.04)
Observed covariate matching	5.74	5.86
	(0.05)	(0.05)

Notes—Means and standard deviations from 250 simulations. In Study 2, the X are drawn independently of J with mass function $p_x = (.2, .3, .3, .2)$ while the finite-mixture model allows for dependence (as in the main text). In Study 3, X and J are dependent as in the main text, but the finite-mixture model is misspecified to assume that they are independent and that the errors in the selection equations are logistic. In Study 4, the data are drawn as in the main text, but the finite-mixture is misspecified to use two, rather than three, latent classes (the results for this study are based on 100 simulations).

Table 3: The effects of gang membership on violent behavior

		Shot	Rob	Beat	Fight
	Naive	0.02	0.07	0.12	0.28
		(0.12)	(0.14)	(0.14)	(0.13)
Population ATT	Uncorr'd	0.02	0.06	0.09	0.23
		(0.11)	(0.14)	(0.11)	(0.15)
	Corr'd	0.02	0.05	0.08	0.22
		(0.14)	(0.12)	(0.22)	(0.14)
	Corr'd - Uncorr'd	0	0	0	0.01
		(0.12)	(0.14)	(0.14)	(0.15)
Population ATE	Uncorr'd	0.02	0.06	0.09	0.22
		(0.14)	(0.12)	(0.2)	(0.12)
	Corr'd	0.03	0.06	0.1	0.23
		(0.15)	(0.11)	(0.14)	(0.12)
	Corr'd - Uncorr'd	0	0	0	0.01
		(0.15)	(0.14)	(0.12)	(0.16)
Group ATEs	1, Uncorr'd	0.06	0.11	0.13	0.24
		(0.15)	(0.12)	(0.11)	(0.12)
	2, Uncorr'd	0	0	0.06	0.19
		(0.14)	(0.13)	(0.11)	(0.1)
	3, Uncorr'd	0	0.04	0.08	0.25
		(0.19)	(0.17)	(0.11)	(0.17)
	1, Corr'd	0.09	0.17	0.19	0.37
		(0.12)	(0.15)	(0.1)	(0.15)
	2, Corr'd	-0.03	-0.06	0.01	0.09
		(0.11)	(0.19)	(0.13)	(0.13)
	3, Corr'd	-0.02	0	0.03	0.15
		(0.15)	(0.14)	(0.14)	(0.15)
	1, Corr'd - Uncorr'd	0.03	0.06	0.06	0.13
		(0.15)	(0.12)	(0.14)	(0.13)
	2, Corr'd - Uncorr'd	-0.03	-0.06	-0.05	-0.1
		(0.15)	(0.19)	(0.11)	(0.17)
	3, Corr'd - Uncorr'd	-0.02	-0.04	-0.05	-0.1
		(0.14)	(0.13)	(0.12)	(0.13)

Notes—Standard errors based on 500 nonparametric bootstrap replications. "Naive" denotes a simple comparison of means between treated and untreated outcomes, "Uncorr'd" denotes the uncorrected latent-class matching estimator that compares treated and untreated units assigned to the same group, and "Corr'd" denotes a latent-class matching estimator that corrects for misclassifications arising in the first stage. Both latent-class matching estimators are based on model assignment. The estimated group-specific propensity scores are .34 (.13), .35 (.13), and .32 (.13) for members of groups one, two, and three (respectively, standard errors in parentheses).