Hybridizing Machine Learning Methods and Finite Mixture Models for Estimating Heterogeneous Treatment Effects in Latent Classes

Youmi Suk [*1], Jee-Seon Kim [†1], and Hyunseung Kang [‡2]

[1]Department of Educational Psychology, University of Wisconsin-Madison
[2]Department of Statistics, University of Wisconsin-Madison

August 28, 2020

Abstract

There has been increasing interest in exploring heterogeneous treatment effects using machine learning (ML) methods such as Causal Forests, Bayesian Additive Regression Trees (BART), and Targeted Maximum Likelihood Estimation (TMLE). However, there is little work on applying these methods to estimate treatment effects in latent classes defined by well-established finite mixture/latent class models. This paper proposes a hybrid method, a combination of finite mixture modeling and ML methods from causal inference to discover effect heterogeneity in latent classes. Our simulation study reveals that hybrid ML methods produced more precise and accurate estimates of treatment effects in latent classes. We also use hybrid ML methods to estimate the differential effects of private lessons across latent classes from TIMSS data.

Keywords: Causal inference, Machine learning methods, Observational studies, Multilevel propensity score matching, Finite mixture modeling

1 Introduction

1.1 Motivation

There has been a growing interest in causal inference to estimate conditional average treatment effects (CATEs) using machine learning (ML) methods (Su, Tsai, Wang, Nickerson, & Li, 2009; Imai & Ratkovic, 2013; Athey & Imbens, 2016; Hill, 2011; Wager & Athey, 2018; Künzel, Sekhon, Bickel, & Yu, 2019). These methods show great promise in understanding treatment effect heterogeneity based on observable characteristics of the study population. However, in some settings, observable characteristics are thought to emerge from meaningful latent processes. For example, many studies in education and psychology posit the existence of latent classes defined by parameters in latent class/finite mixture models in order to better understand observed student behaviors such as internet and smartphone addiction or teen smoking (Clogg, 1995; McLachlan & Peel, 2000; Mok et al., 2014; Sutfin, Reboussin, McCoy, & Wolfson, 2009). Differences in observed behaviors are hypothesized to arise due to differences in latent classes, and, as such, there is a strong emphasis on understanding the differences between latent classes by examining the parameters of latent class/profile models or latent class regression models; see Magidson and Vermunt (2004) for details. In such cases where latent classes play a vital role in the scientific understanding of

[*]ysuk@wisc.edu
[†]jeeseonkim@wisc.edu
[‡]hyunseung@stat.wisc.edu

This article has been accepted for publication in *Journal of Educational and Behavioral Statistics*, published by SAGE Publishing.

observed phenomena, understanding how the effects of a new treatment, program, or policy vary across these latent classes is of great interest.

To provide a concrete example that motivated this work, consider an observational study estimating the effect of taking private science lessons (i.e., treatment) on science test scores (i.e., outcome) among middle school students. Each student's choice to have private tutors is based on a number of observable characteristics, such as their previous grades and the location of their schools, as well as unobservable, latent characteristics, such as students' motivations, academic resilience[1] , or science self-efficacy[2]. For example, some students who are academically resilient may seek private tutors to supplement classroom instruction compared to those who are less resilient. Some students may opt for a private tutor because they are self-motivated, while others may not seek a tutor because they are less motivated. Or, the driving factor for private lessons may be similar for all students in the same school because of deficiencies (or lack thereof) in school resources. Regardless, these characteristics may not be directly observable, but rich latent class models exist in psychology to help us better understand them; see McLachlan and Peel (2000), Kaplan, Kim, and Kim (2009), and Masyn (2013) for examples. More importantly, variation in these latent classes may lead to differential effects of having a private tutor. For instance, a private tutor may be more helpful in raising test scores among students who are academically resilient or self-motivated compared to students who are less resilient or less motivated.

If an investigator uses one of the aforementioned ML-based estimator to study effect heterogeneity of private tutoring, these methods will only reveal variations in treatment effects among observable characteristics of the student; they would not be able to reveal variations in treatment effects among latent classes representing resilience and self-motivation. To better illustrate this point, consider Figure 1, where we constructed a hypothetical two-class latent structure and students belong to either one of the two latent classes; say one class represents strong academic resilience, while another class represents weak academic resilience. The average treatment effect of having private tutors in the first latent class (in yellow) is two, whereas the average treatment effect in the second latent class (in green) is zero. When we use Causal Forests (Athey, Tibshirani, & Wager, 2019), an ML-based causal inference method based on random forests, to estimate treatment effects, the Causal Forests masks these two latent classes' treatment effects. In contrast, our hybrid methods, which we explain below, are able to reveal the two latent classes and their respective treatment effects. Specifically, Figure 1 uses our hybrid methods based on Causal Forests, which we refer to as Hybrid Causal Forests.

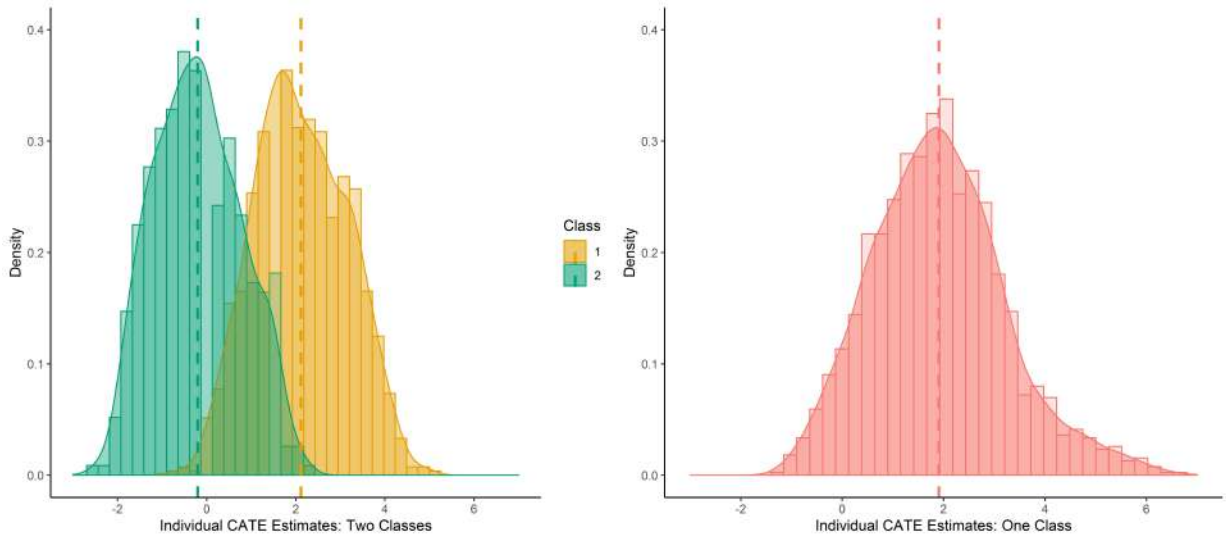

Figure 1: Distributions of individual CATE estimates. The left plot shows our method, Hybrid Causal Forests, while the right plot shows Causal Forests. Dashed lines represent class-specific treatment effect estimates; see Section 4 below for details.

1.2 Prior Work and Our Contribution

Prior works on treatment effect estimation in latent classes are diverse. Kang and Schafer (2010) and Schuler, Leoutsakos, and Stuart (2014) used latent class models to identify latent treatment classes based

on manifest/observed items. Butera, Lanza, and Coffman (2014) and Lanza, Coffman, and Xu (2013) discussed estimating treatment effects when the outcome variables are latent classes. Specifically, the underlying construct of the outcomes was measured using observed items and the goal was to estimate treatment effects on latent class membership of the outcomes. Jo, Wang, and Ialongo (2009) discussed latent trajectory structures in three outcome measures of attention deficit among children and revealed heterogeneity in longitudinal outcomes across latent classes. The work most related to ours is by Kim and Steiner (2015) who used a latent class regression model to model different latent representations of students' selection into treatment (or control) and used multilevel propensity score matching to estimate the treatment effect within each latent class.

The goal of this paper is to complement these prior works and provide a general "hybrid" framework to study treatment effect variation within latent classes by combining latent class modeling with ML-based methods in causal inference. Specifically, in a two-level setting common in education, we propose a two-step hybrid procedure that first uses latent class/finite mixture modeling to identify latent class structures and second, uses modern ML methods in causal inference to estimate treatment effects within each latent class. Our rationale for using ML methods in the second step is to leverage ML's flexibility in modeling potentially complex propensity score and outcome regression models in each latent class. More broadly, our approach to this problem follows a growing trend of combining ML methods with well-established models in psychology to capitalize on the advantages of each approach (Ma, 2018; Suk, Kang, & Kim, 2019). The paper focuses on three popular ML methods in causal inference—Causal Forests (Wager & Athey, 2018; Athey et al., 2019), Bayesian additive regression trees (BART) (Hill, 2011), and Targeted Maximum Likelihood Estimation (TMLE) (Van Der Laan & Rubin, 2006)— but our framework can be extended to other ML methods. We validate our proposed methods through a simulation study and a large-scale educational assessment study concerning the effect of private science lessons on science achievement scores. We show that our ML-based hybrid methods have more precise and accurate estimates of the variations in treatment effects associated with latent classes than other parametric methods used in this research.

2 Review: Notation, Causal Assumptions, and the Propensity Score

We use the Neyman-Rubin causal model (Rubin, 1974; Neyman, 1923) and its extension to multilevel data to define causal effects (Hong & Raudenbush, 2006). Let $Y_{ij}(1)$ be the potential outcome if individual i at cluster j were to be treated ($Z_{ij} = 1$). Let $Y_{ij}(0)$ be the potential outcome if individual i at cluster j were to be untreated ($Z_{ij} = 0$). The notation assumes the stable unit treatment value assumption (SUTVA; Rubin, 1986) where the potential outcomes of each individual are not affected by others' treatment assignments and there is only a single version of treatment. This allows us to write the observed outcome Y_{ij} as $Y_{ij} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$. Let $\mathbf{X}_{ij}$ and $\mathbf{W}_j$ denote pre-treatment covariates for individual i in cluster j, where $\mathbf{X}_{ij}$ are individual-specific covariates and $\mathbf{W}_j$ are cluster-specific covariates.

Under the potential outcomes framework, we assume *strong ignorability*:

$$Y_{ij}(1), Y_{ij}(0) \perp Z_{ij}|\mathbf{X}_{ij}, \mathbf{W}_j \quad \text{and} \quad 0 < e(\mathbf{X}_{ij}, \mathbf{W}_j) = Pr(Z_{ij} = 1|\mathbf{X}_{ij}, \mathbf{W}_j) < 1$$

where $\perp$ denotes independence between two random variables and $e(\mathbf{X}_{ij}, \mathbf{W}_j)$ is the propensity score (Rosenbaum & Rubin, 1983). In single-level data, propensity scores are typically estimated with logistic regression. In multilevel data, propensity scores are typically estimated with random or fixed effects logistic regression (Leite, 2016). Propensity scores are often used in matching methods to match treated and control units or to weigh individuals' outcomes via inverse probability weighing.

We conclude the section by defining the causal estimand of interest. Let $K_{ij} = \{1, \ldots, C\}$ denote the latent class membership of individual i in cluster j from a latent class model. The goal in this paper is to estimate the CATE for individuals who belong to latent class $K_{ij} = k$ and is formalized as follows:

$$\tau(k) = E[Y_{ij}(1) - Y_{ij}(0) \mid K_{ij} = k]$$

The estimand $\tau(k)$ cannot be directly estimated with observed data since latent class membership K_{ij} is unobserved. More precisely, the function that relates the observable characteristics $\mathbf{X}_{ij}, \mathbf{W}_j$ to their latent counterparts K_{ij} is unknown and must be modeled based on context-specific finite mixture/latent class models. In contrast, the usual CATE formalized below

$$\tau(\mathbf{x}, \mathbf{w}) = E[Y_{ij}(1) - Y_{ij}(0) \mid \mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w}]$$

can be directly estimated from the observed data under strong ignorability and SUTVA; see Imbens and Rubin (2015) for more details about identification of CATE. Modern ML methods in causal inference (e.g., Causal Forests) provide consistent and asymptotically Normal estimates of $\tau(\mathbf{x}, \mathbf{w})$. However, $\tau(\mathbf{x}, \mathbf{w})$ only reveals treatment heterogeneity among observable characteristics and masks treatment variability in latent classes. The next section discusses our proposed approach, hybrid ML methods, which sequentially integrate latent class modeling and ML methods to estimate $\tau(k)$.

3 Hybridizing Latent Class Modeling and Machine Learning for Causal Inference

Our hybrid approach has two steps. The first step estimates latent classes via context-specific latent class/finite mixture modeling, and the second part uses ML-based methods to estimate treatment effects within each latent class. Subsequent sections elaborate on each step.

3.1 Step 1: Latent Class/Finite Mixture Modeling

Latent class or finite mixture models have been frequently used to group individuals or data into unobserved latent classes that can be inferred from the observed data (McLachlan & Peel, 2000; Vermunt & Magidson, 2003). In a standard latent class model, latent classes are identified using categorical latent class indicators e.g., dichotomous survey items, and parameters defining latent classes are response probabilities (Muthén & Muthén, 2017; Wang & Wang, 2012). More generally, there are many types of latent class/finite mixture models based on regression analysis, path analysis, and factor analysis; see Magidson and Vermunt (2004), McLachlan and Peel (2000), Kaplan et al. (2009), and Masyn (2013) for more details. The choice of which latent model to use is context-specific and so long as researchers choose an identifiable finite mixture model to estimate latent classes and each class meets the aforementioned casual assumptions, our methodology will work.

In our real data study of private science tutoring and school achievement scores, we focus on a type of latent class models that describe how students in each school select themselves into treatment (i.e., private science tutoring), also referred to as latent selection/propensity score models, and latent class membership applies at the cluster level (i.e., at the school level) so that $K_{ij} = k$ for everyone in the same cluster. In particular, each cluster belongs to one of $k = 1, \ldots, C$ propensity score models that govern how students within each school select themselves into treatment, and the parameters of each propensity score define the latent classes; in other words, students' selection behaviors in private tutoring are homogeneous within each cluster, but there are hidden heterogeneous structures across schools that can be inferred from the data. As an illustration of the chosen latent class model, suppose that school principals emphasize academic achievements. In such schools, students may seek private lessons to receive higher achievement scores. Additionally, although a school principal's emphasis on academic achievement can play a role in students seeking private tutors, its importance may differ depending on the location of the school; private education services are more readily available in urban areas than in rural areas. We remark that this model is also called a restricted multiple group latent class model (Vermunt, 2003) and Kim and Steiner (2015) provides additional interpretations as well as some limitations of the model.

The overall goal of the latent class selection model in our real data example is to understand which of the C selection models govern students' choices to select private tutors. Formally, let $\pi_k = P(K_j = k), k = 1, \ldots, C$ be the marginal probability of being in latent class k. For this latent class model, we drop the subscript i in K_{ij} for clarity, but we can define $K_{ij} = K_j$ to fit it into the general notation and we use the two notations interchangeably. Consider a latent class random effects logistic regression model where each student's choice to seek treatment (e.g., private tutors) is a mixture of C different selection models. Specifically, each latent class k has its own selection model $e_k(\mathbf{X}_{ij}, \mathbf{W}_j, \theta_k) = P(Z_{ij} = 1 \mid \mathbf{X}_{ij}, \mathbf{W}_j, \theta_k)$ where θ_k parameterizes the model; for two-level data, the selection model for each latent class k is a random effects logistic model. Then, a latent class selection model assumes that $P(Z_{ij} = 1 \mid \mathbf{X}_{ij}, \mathbf{W}_j)$ is C mixtures of selection models with mixing probabilities π_k, $k = 1, \ldots, C$:

$$P(Z_{ij} = 1 \mid \mathbf{X}_{ij}, \mathbf{W}_j) = \sum_{k=1}^{C} \pi_k e_k(\mathbf{X}_{ij}, \mathbf{W}_j, \theta_k), \quad 0 \leq \pi_k \leq 1, \ \sum_{k=1}^{C} \pi_k = 1 \tag{1}$$

where

$$\pi_k = P(K_j = k) = \frac{exp(\gamma_k)}{\sum_{k=1}^{C} exp(\gamma_k)} \tag{2}$$

That is, π_k is modeled by a multinomial logistic model with γ_k representing a class-specific multinomial intercept. The parameters in the models (i.e., $\theta_k, \pi_k, k = 1, \ldots, C$) are estimated by an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Peel, 2000; Leisch, 2004). Using Bayes rule, we can compute the posterior probability that each cluster belongs to latent class k. Specifically, individual ij is assigned to one of C latent classes with the highest posterior probability, also known as modal assignment. Some alternatives to modal assignment are proportional assignment and random assignment (Vermunt, 2010; Goodman, 2007). Here, we use modal assignment due to its simplicity and optimality under certain assumptions about Bayes classification error rates (Bakk, Tekle, & Vermunt, 2013). Regardless, let $\hat{K}_j$ (or $\hat{K}_{ij} = \hat{K}_j$) denote the estimated latent class membership for each cluster. We will use the estimated membership in the second step of our proposed algorithm.

There are some important implementation details in estimating latent class models and we briefly summarize four issues that are most relevant to our setting; see Everitt and Hand (1981), Titterington, Smith, and Makov (1985), and McLachlan and Peel (2000) for detailed discussions. First, typically, the number of latent classes C is initially specified based on subject-matter theories about latent class structure, but later verified by a data-driven approach based on various measures of model fit, such as the likelihood ratio statistic, Pearson's Chi-square, the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (Kaplan et al., 2009). While there is always a risk of incorrectly specifying the number of latent classes, either through over-extraction (i.e., more latent classes were specified than the true number of latent classes) or under-extraction (i.e., fewer latent classes were specified than the true number of latent classes), it is generally preferable to have more latent classes than fewer latent classes for estimating treatment effects, as the former would be able to identify diverse structures of latent classes. Second, latent class models are only identifiable up to labeling permutations of latent classes because estimated class labels are arbitrary. For example, if there are three latent classes, the estimated parameters for Class 1 can equally be labeled as Class 2 or Class 3 and the data will be observationally equivalent. This issue primarily affects label assignment, but not estimation of model parameters (Leisch, 2004). Also, many algorithms have been developed to detect label switching issues and to relabel latent classes using ordering constraints and Stephens' methods (e.g., Tueller, Drotar, & Lubke, 2011; Stephens, 2000). Third, in general, mixtures of univariate Normal, gamma, exponential, Cauchy, and Poisson distributions are identifiable, whereas mixtures of uniform distributions are not identifiable. Mixtures of binomial and multinomial distributions can be identified under some assumptions on the number of latent classes and the size of the support of $Z_{ij}, \mathbf{X}_{ij}$, and $\mathbf{W}_j$ (Grün & Leisch, 2008; Everitt & Hand, 1981; Titterington et al., 1985; Allman, Matias, Rhodes, et al., 2009). Finally, to prevent over-fitting the latent selection model, we can set the prior class probabilities to be far away from zero and set θ_k to be sufficiently different (Leisch, 2004). Adding random effects in e_k can also help avoid overfitting (Lenk & DeSarbo, 2000).

For software to implement step 1, we used the software *Mplus8* (Muthén & Muthén, 2017) and the MplusAutomation package (Hallquist & Wiley, 2017) in R (R Core Team, 2019) to estimate the latent class $\widehat{K}_{ij}$. We also applied a "class assignment based algorithm" to resolve potential label switching issues (Tueller et al., 2011).

3.2 Step 2: Machine Learning Methods for Causal Inference

This section describes some methods in causal inference that utilize ML to estimate heterogeneous treatment effects. We remind readers that the specific choice of the ML method is not critical so long as they provide consistent point estimators and valid confidence intervals. Also, for one of the ML methods, Causal Forests, we follow a suggestion from recent work (Suk et al., 2019) to improve its performance in two-level data.

3.2.1 General Approach

At a high level, almost all ML-based methods for causal inference require estimating either the outcome model, the propensity score model, or both. Briefly, the outcome model is the conditional expectation of the outcome given the observed covariates and treatment assignment, say $m(\mathbf{x}, \mathbf{w}, z) = E[Y_{ij} \mid \mathbf{X}_{ij} =$

$\mathbf{x}, \mathbf{W}_j = \mathbf{w}, Z_{ij} = z]$; some methods also estimate the conditional expectation of the outcome given only the covariates, say $m(\mathbf{x}, \mathbf{w}) = E[Y_{ij} \mid \mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w}]$. As mentioned above, the propensity score model is the probability of being assigned to treatment given observed covaraites, i.e., $e(\mathbf{x}, \mathbf{w}) = P(Z_{ij} = 1 \mid \mathbf{X}_{ij} = \mathbf{x}, \mathbf{W}_j = \mathbf{w})$. Each ML-based method in causal inference estimates the outcome model or the propensity score model using different supervised ML algorithms. For example, BART uses Bayesian regression trees to estimate $m(\cdot)$. TMLE, combined with SuperLearner (van der Laan, Polley, & Hubbard, 2007), uses an ensemble of supervised learning algorithms to estimate $m(\cdot)$ and $e(\cdot)$. Causal Forests uses a modified random forest to estimate $m(\cdot)$ and $e(\cdot)$. Also, each ML-based method aggregates estimates of $m(\cdot)$ and $e(\cdot)$ differently to arrive at the final estimate of CATE. For example, typical BART only uses $m(\cdot)$ to estimate CATE. Causal Forests, which we describe below in Section 3.2.2, uses $m(\cdot)$ and $e(\cdot)$ through a weighted linear regression approach. TMLE uses both $m(\cdot)$ and $e(\cdot)$ through a "clever covariate" to estimate the CATE. If the underlying supervised ML algorithm can consistently estimate $m(\cdot)$ and $e(\cdot)$, these methods not only provide a consistent estimate of the CATE but also, under additional assumptions, provide valid p-values and confidence intervals.

To incorporate ML-based methods into latent class estimation in two-level settings, we outline the following approach. First, for each estimated latent class k, use any of the aforementioned ML-based CATE estimators to estimate the CATE within each k by only using the data from the latent class and denote this as $\tau(\mathbf{x}, \mathbf{w}, k)$; note that if the encompassing ML method requires estimation of $e(\cdot)$, one can use a random effects logistic regression model instead of the associated superivsed learning algorithm to improve performance in clustered/multilevel data. Second, average $\tau(\mathbf{x}, \mathbf{w}, k)$ among individuals with the same k to arrive at the final estimator for $\tau(k)$. We show an example of this general recipe based on Causal Forests below.

3.2.2 Vanilla Causal Forests and Modified Causal Forests

Causal Forests (Wager & Athey, 2018; Athey et al., 2019) is a type of random forests (Breiman, 2001) that is used to estimate the CATE as well as the average treatment effect. Specifically, a Causal Forest estimator of the CATE is a weighted linear regression of residualized outcome $\tilde{Y}_{ij} = Y_{ij} - \hat{m}^{(-ij)}(\mathbf{x}, \mathbf{w})$ and a single residualized regressor $\tilde{Z}_{ij} = Z_{ij} - \hat{e}^{(-ij)}(\mathbf{x}, \mathbf{w})$.

$$\hat{\tau}(\mathbf{x}, \mathbf{w}) = \frac{\sum_{ij} \alpha_{ij}(\mathbf{x}, \mathbf{w})(Y_{ij} - \hat{m}^{(-ij)}(\mathbf{X}_{ij}, \mathbf{W}_j))(Z_{ij} - \hat{e}^{(-ij)}(\mathbf{X}_{ij}, \mathbf{W}_j))}{\sum_{ij} \alpha_{ij}(\mathbf{x}, \mathbf{w})(Z_{ij} - \hat{e}^{(-ij)}(\mathbf{X}_{ij}, \mathbf{W}_j))^2} \tag{3}$$

Here, $0 \leq \alpha_{ij}(\mathbf{x}, \mathbf{w}) \leq 1$ weighs how much each unit ij contributes to the estimate of CATE, $\tau(\mathbf{x}, \mathbf{w})$. The $^{(-ij)}$-superscript represents out-of-bag leave-one-out estimates in machine learning, i.e., the estimates of functions when unit ij's data is not used. In Causal Forests, the estimates of $\hat{m}^{(-ij)}(\mathbf{x}, \mathbf{w})$ and $\hat{e}^{(-ij)}(\mathbf{x}, \mathbf{w})$ are obtained by a honest random forest algorithm in Procedure 1 of Wager and Athey (2018). Wager and Athey (2018) and Athey et al. (2019) showed that the Causal Forests estimator is consistent for the CATE and has an asymptotic pivotal Gaussian distribution under some assumptions; the latter property allows researchers to construct valid p-values and confidence intervals for the CATE.

To estimate class-specific treatment effects in two-level data using Causal Forests, we do the following. First, instead of using a random forest to estimate the propensity score, we use a multilevel logistic regression in step 1 to account for clustering structures inside Causal Forests (Suk et al., 2019). Second, we run Causal Forests among units that are in the same latent class k. Combined, the modified CATE estimator using Causal Forests can be formalized as:

$$\hat{\tau}(\mathbf{x}, \mathbf{w}, k) = \frac{\sum_{ij:\widehat{K}_{ij}=k} \alpha_{ij}(\mathbf{x}, \mathbf{w})(Y_{ij} - \hat{m}_k^{(-ij)}(\mathbf{X}_{ij}, \mathbf{W}_j))(Z_{ij} - \hat{e}_k(\mathbf{X}_{ij}, \mathbf{W}_j))}{\sum_{ij:\widehat{K}_{ij}=k} \alpha_{ij}(\mathbf{x}, \mathbf{w})(Z_{ij} - \hat{e}_k(\mathbf{X}_{ij}, \mathbf{W}_j))}$$

Note that $\hat{m}_k^{(-ij)}$ bears a subscript k to denote that it has been estimated using data from individuals who belong to latent class k. Also, $\hat{e}_k$ no longer has the $^{(-ij)}$-superscript to denote that it has been estimated using a multilevel logistic regression instead of the default regression forests. Averaging $\hat{\tau}(\mathbf{x}, \mathbf{w}, k)$ across all individuals in the same latent class k is our estimate of the average treatment effect within latent class k, i.e.,

$$\hat{\tau}(k) = \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \hat{\tau}(\mathbf{x}, \mathbf{w}, k)$$

where N_k denotes the sample size in each latent class, k.

Finally, we briefly remark that instead of the proposed approach, an alternative approach to combine latent class estimates with ML methods is to use the estimated latent class as a "covariate" in ML methods; see Appendix A. We show in the Appendix that our approach has better finite sample performance than the alternative approach in terms of bias and mean squared error (MSE).

4 Simulation Study

4.1 Simulation Design and Evaluation

We conducted a simulation study to investigate the performance of hybrid ML methods. Our data generating models follow Kim and Steiner (2015), Kim, Steiner, and Lim (2016), and our motivating data, which had a two-level structure with one continuous outcome and one binary treatment. We consider four continuous covariates, two of which are individual-level covariates and the other two are cluster-level covariates. We also assume two latent classes defined by the latent selection model discussed before and each latent class has its own unique treatment effect. The details of our data generating procedure are stated below.

1. Let $nC1$ and $nC2$ represent the number of clusters in latent class $k = 1$ and latent class $k = 2$. For each cluster in each latent class, we generate the number of individuals n_j based on drawing samples from a Normal distribution with mean nS set to either 30 or 50, variance v set to either 4 or 16, and round them to the nearest integer.
2. For each individual i in cluster j, randomly sample two cluster-level covariates, $\mathbf{W}_j = (W_{1j}, W_{2j})$ and two individual-level covariates, $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij})$, from the following distributions

$$\begin{pmatrix} W_{1j} \\ W_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & .2 \\ .2 & 2 \end{pmatrix}\right], \quad \begin{pmatrix} X_{1ij} \\ X_{2ij} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix}, \begin{pmatrix} 10 & 2 \\ 2 & 15 \end{pmatrix}\right]$$
$$\begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix} = \begin{pmatrix} .1 & .05 \\ .08 & .1 \end{pmatrix} \begin{pmatrix} W_{1j} \\ W_{2j} \end{pmatrix} + \begin{pmatrix} \kappa_{1j} \\ \kappa_{2j} \end{pmatrix}, \quad \begin{pmatrix} \kappa_{1j} \\ \kappa_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}\right]$$

   We remark that the individual-level covariates' means μ_j are a function of the cluster-level covariates $\mathbf{W}_j$ and random errors κ_j. We set larger variances for individual-level covariates than cluster-level covariates to reflect typically higher variations in individual-level covariates than cluster-level covariates.
3. For each latent class $k = 1, 2$, define the propensity score model $e_k(\cdot)$ based on a random effects logistic selection model.

$$\begin{aligned} logit(e_1(\mathbf{X}_{ij}, \mathbf{W}_j)) &= 0 + 0.15X_{1ij} + 0.1X_{2ij} + 0.1W_{1j} + 0.2W_{2j} + 0.1X_{1ij}W_{2j} + R_{j1} \\ logit(e_2(\mathbf{X}_{ij}, \mathbf{W}_j)) &= -0.05 + 0.05X_{1ij} - 0.05X_{2ij} + 0.2W_{1j} + 0.05W_{2j} + R_{j2} \\ R_{j1} \sim N(0, 0.5), &\quad R_{j2} \sim N(0, 0.2) \end{aligned}$$

   Here, $e_k(\cdot)$ is the propensity score for individual i in cluster j which belongs to latent class k. R_{jk} is the random effect for cluster j in class k. The slope coefficients for Class 1 and Class 2 differ where Class 1 has a stronger selection than Class 2. The intra-class correlations for Class 1 and Class 2 are around 0.13 and 0.06, respectively.
4. For each individual i in cluster j which belongs to latent class k, generate individual treatment status, Z_{ij} (0 = untreated; 1=treated) from a Bernoulli distribution with the propensity score specified above.

$$Z_{ij} \sim \begin{cases} Bernoulli(e_1(\mathbf{X}_{ij}, \mathbf{W}_j)), & \text{if } i \text{ belongs to latent class } k = 1 \\ Bernoulli(e_2(\mathbf{X}_{ij}, \mathbf{W}_j)), & \text{if } i \text{ belongs to latent class } k = 2 \end{cases}$$

5. For each individual i in cluster j which belongs to latent class k, generate potential outcomes and observed outcomes based on random effects linear regression models.

$$\begin{aligned} Y_{ij1}(z) &= 100 + 2.5 \cdot z + 2X_{1ij} + 1X_{2ij} + 2W_{1j} + 1.5W_{2j} + 0.5X_{2ij}W_{1j} + 0.3X_{2ij}^2 + U_{j1} + \epsilon_{ij1} \\ Y_{ij2}(z) &= 80 + 0 \cdot z + 1X_{1ij} + 0.5X_{2ij} + 1W_{1j} + 0.5W_{2j} + 0.2X_{2ij}W_{1j} + 0.2W_{1j}W_{2j} + U_{j2} + \epsilon_{ij2} \\ Y_{ij} &= Z_{ij}Y_{ijk}(1) + (1 - Z_{ij})Y_{ijk}(0), \quad U_{j1} \sim N(0, 10), \quad U_{j2} \sim N(0, 7), \quad \epsilon_{ijk} \sim N(0, 100) \end{aligned}$$

The term U_{jk} is the random effect for cluster j in latent class k, and ϵ_{ijk} is the random error for individual i in cluster j which belongs to latent class k. The treatment effect is positive for Class 1, but zero for Class 2 so that each latent class has distinct treatment effects. The intra-class correlations are 0.10 and 0.07 for Classes 1 and 2, respectively. Additionally, there are non-linear and/or interaction terms in the outcome model.

In our simulation study, we varied the following simulation parameters: the size of each latent class, $nC1$ and $nC2$, and the mean size of each cluster nS. We examined the performance of hybrid ML methods—Hybrid Causal Forests, Hybrid BART, and Hybrid TMLE—in estimating latent class average treatment effects $\tau(k)$. In Appendix D, we examined the performance of the estimated individual CATE. As for software, we use the grf package (Tibshirani et al., 2019) for Causal Forests, bartCause package (Dorie & Hill, 2019) for BART, and tmle package (Gruber & van der Laan, 2012) for TMLE, all implemented in R (R Core Team, 2019) As a comparison, we ran within-class matching (Kim & Steiner, 2015; Kim et al., 2016) as an alternative to hybrid ML methods which estimate average treatment effects within each latent class via propensity score within-class matching. In brief, within-class matching is a type of multilevel matching that matches treated and control units across clusters, but within the same latent classes defined by latent selection models. Within-class matching uses the same latent selection model as above to identify latent classes and requires specifying a weighing function that depends on the estimated propensity score to obtain estimates of the treatment effect within each latent class. For our simulation, the propensity score for within-class matching was estimated using random effects logistic regression. For weighing, we used inverse probability weighting (IPW) and marginal mean weighing through stratification (MMW-S) (Hong & Hong, 2009). Specifically, the IPW estimator for latent class k using within-class matching is

$$\hat{\tau}_{\text{IPW}}(k) = \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \frac{Y_{ij}Z_{ij}}{e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} - \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \frac{Y_{ij}(1-Z_{ij})}{1-e_k(\mathbf{X}_{ij}, \mathbf{W}_j)}$$

and the MMW-S estimator for latent class k using within-class matching is

$$\omega_{z,ij(k)} = \begin{cases} \frac{E_{z1(k)}}{O_{z1(k)}} & \text{if } e_k(\mathbf{X}_{ij}, \mathbf{W}_j) \text{ in stratum 1 of latent class } k \\ \vdots & \\ \frac{E_{zS(k)}}{O_{zS(k)}} & \text{if } e_k(\mathbf{X}_{ij}, \mathbf{W}_j) \text{ in stratum } S \text{ of latent class } k \end{cases}$$

$$\hat{\tau}_{\text{MMW-S}}(k) = \frac{1}{N} \sum_{ij:\widehat{K}_{ij}=k} Y_{ij}Z_{ij}\omega_{1,ij(k)} - \frac{1}{N} \sum_{ij:\widehat{K}_{ij}=k} Y_{ij}(1-Z_{ij})\omega_{0,ij(k)}$$

where $O_{zs(k)}$ is the observed frequency of individuals in treatment status $z \in \{0,1\}$ and stratum $s \in \{1,2,...,S\}$ of the distribution of the propensity score in latent class k, and $E_{zs(k)}$ is the expected frequency assuming the distributions between treated and untreated units are the same across strata. We created 10 strata of propensity scores for MMW-S. We also computed the doubly robust (DR) estimator as follows:

$$\begin{aligned} \hat{\tau}_{DR}(k) = \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \left[\frac{Z_{ij}Y_{ij}}{e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} - \frac{\{Z_{ij} - e_k(\mathbf{X}_{ij}, \mathbf{W}_j)\}}{e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} m_k(\mathbf{X}_{ij}, \mathbf{W}_j, 1) \right] \\ - \frac{1}{N_k} \sum_{ij:\widehat{K}_{ij}=k} \left[\frac{(1-Z_{ij})Y_{ij}}{1-e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} + \frac{\{Z_{ij} - e_k(\mathbf{X}_{ij}, \mathbf{W}_j)\}}{1-e_k(\mathbf{X}_{ij}, \mathbf{W}_j)} m_k(\mathbf{X}_{ij}, \mathbf{W}_j, 0) \right] \end{aligned}$$

Here, $e_k(\mathbf{X}_{ij}, \mathbf{W}_j)$ is the propensity score estimated by random-effects logistic regression models within each latent class k. $m_k(\mathbf{X}_{ij}, \mathbf{W}_j, 0)$ and $m_k(\mathbf{X}_{ij}, \mathbf{W}_j, 1)$ are outcome models based on random-effects linear regression models within each latent class k. Both models are specified to be the same as those from the data generating models. For more details on other propensity score techniques, see Schafer and Kang (2008), Austin (2011), and Steiner and Cook (2013).

Each method was evaluated based on the absolute bias and MSE of class-specific average treatment effect estimates. Specifically, given $m = 1, \ldots, 400$ simulation replications and their corresponding estimates $\hat{\tau}_m(k)$ $(m = 1, \ldots, 400)$, the absolute bias and MSE within each class are defined as:

$$|\text{Bias}(k)| = \frac{1}{400} \Big| \sum_{m=1}^{400} (\hat{\tau}_m(k) - \tau(k)) \Big|, \quad \text{MSE}(k) = \frac{1}{400} \sum_{m=1}^{400} (\hat{\tau}_m(k) - \tau(k))^2$$

We also evaluate the overall performance across latent classes by computing the overall bias and MSE as follows:

$$|\text{Bias}| = \frac{1}{400}\Big|\sum_{m=1}^{400}\sum_{k=1}^{2}\frac{N_k^*}{N_m}(\hat{\tau}_m(k)-\tau(k))\Big|, \quad \text{MSE} = \frac{1}{400}\sum_{m=1}^{400}\sum_{k=1}^{2}\frac{N_k^*}{N_m}(\hat{\tau}_m(k)-\tau(k))^2$$

The term N_k^* denotes the true sample size in each latent class k and N_m denotes the total sample size in each simulation replication.

4.2 Simulation Results

Table 1 summarizes the mean percentage of correctly identifying latent class membership. Mean percentages for modal assignment were calculated by comparing the true class membership of each individual with the estimated class membership from the mixture model, while for proportional assignment, the mean percentages were computed by using a weighted average of the latent class posterior probabilities in each true class. We found that modal assignment classified the latent classes more accurately than proportional assignment. Also, classification rates were affected by cluster sizes and the number of clusters. In particular, we found that increasing the size of the clusters had a larger impact on classification rates than increasing the number of clusters.

Table 1: Classification rate (%) in class membership

(nC1, nC2, nS)	Modal Assignment	Proportional Assignment
(25, 25, 30)	73.42	71.04
(25, 25, 50)	84.04	81.22
(50, 50, 30)	78.81	74.92

Note: nC1, nC2, and nS represent the number of clusters for the first latent class, the number of clusters for the second latent class, and average cluster sizes, respectively.

Figure 2 displays results of class-specific average treatment effect estimates; see Table 5 in Appendix B for numerical results. Across simulation conditions, hybrid ML methods generally performed better than DR and non-DR methods in terms of overall bias and MSE. Indeed, it is not surprising that with the misclassification of the latent classes, the outcome model and/or the propensity scores are inherently incorrect inside the DR and non-DR estimators and thus, traditional parametric methods—IPW, MMW-S, and DR estimators—are directly affected by misclassified units in latent classes. In contrast, ML methods are often "local" non-parametric methods which are more robust to model mis-specifications. Of course, if the misclassification rate is fairly high, then it is unlikely that any method will work properly. Also, as seen from Figure 2, when the sample sizes increased from (25, 25, 30) to (25, 25, 50) or (50, 50, 30), we saw that overall bias and overall MSE decreased across different estimators, but the magnitudes varied depending on the exact sample proportions within each latent class.

We remark that the bias of the IPW estimator in latent class 1 was surprisingly small, but the overall bias was still larger than other methods. This suggests that the IPW estimator traded off a large bias reduction in latent class 1 at the expense of an increase in bias in latent class 2. But this trade-off was not "balanced" and led to a large overall bias. In contrast, hybrid methods like BART had a large bias in latent class 1 and a small bias in latent class 2 but achieved the smallest overall bias. In Appendix C, we further examined the bias trade-off between latent classes and found that the IPW estimator exhibited this phenomenon in other settings. Overall, our results demonstrate that hybrid ML methods provides accurate and precise estimates of the treatment effect and are an attractive alternative to those based on parametric propensity score techniques, IPW and MMW-S, or parametric DR methods.

5 TIMSS Data Study: The Effects of Private Science Lessons

5.1 Data and Variables

We revisit the question in the introduction and study the heterogeneous effects of private science lessons on students' science achievement scores where we suspected distinct latent selection processes across clusters of students. The data comes from the 2015 Trends in International Mathematics and Science

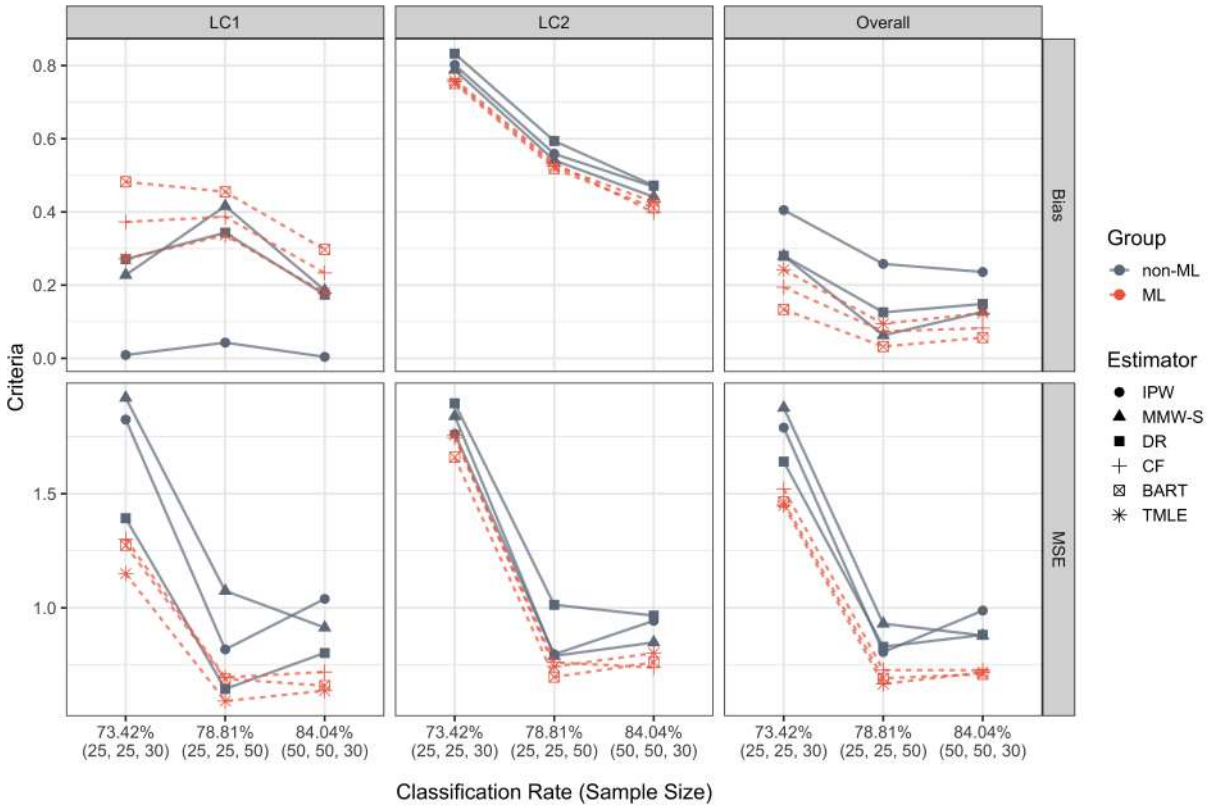

Figure 2: Performance of class-specific treatment effect estimates with classification rates and sample sizes. The three values in parentheses represent the number of clusters for the first latent class, the number of clusters for the second latent class, and the average cluster sizes, respectively. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents the doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation. The true treatment effects are 2.5 and 0 for the first and second latent classes, respectively.

Study (TIMSS) data. TIMSS is an international educational assessment that examines the progression of students' performance in mathematics and science and it was first conducted in 1995 by the International Association for the Evaluation of Educational Achievement (IEA). Since 1995, TIMSS has been conducted for 4-th and 8-th graders every four years in more than 40 countries. The recent data collection, which took place in 2015, was conducted in 60 countries and a new data collection was planned in 2019. The data are based on a two-stage stratified cluster sampling; schools are chosen first according to each country's important demographic variables (e.g., in Korea, school location, and/or whether schools are gendered), and then at least one intact classroom is randomly chosen from each school (Martin, Mullis, & Hooper, 2016). We used the Korea TIMSS 2015 data of 8th graders for our analysis.

The original data included 5,309 students from 150 middle schools with varying school sizes (a range of 6 to 75; mean of 35.4 students per school; median of 32 students per school). We removed students with 1) inconsistent responses about their attendance of private science lessons and 2) missing information in 7 out of 12 covariates (see below for a list of covariates). Our final sample was 4,874 students (91.81% of the original data) from 149 schools. For simplicity and to demonstrate the new methodology, we did not consider multiple plausible values of student achievements in the sciences and ignored sampling weights. However, to rigorously evaluate the effects of private lessons and generalize these results, it is necessary to consider five different plausible values and sampling weights; see Rutkowski, Gonzalez, Joncas, and von Davier (2010) and Foy, Arora, and Stanco (2017) for details.

The treatment variable was whether a student received private science lessons ($Z_{ij} = 1$) or not ($Z_{ij} = 0$). The outcome Y_{ij} was the first plausible value of achievement in the sciences. We included 12 covariates that affected the selection and outcome processes, including six student-level covariates $\mathbf{X}_{ij}$ and six school-level covariates $\mathbf{W}_j$. The student-level covariates were student's gender (*male*), fathers' highest

education levels (*dad.edu*, with three levels; no college, college graduates *dad.cll*, and don't know *dad.q*), the number of books at home (*books25*, with two levels; more than 25, and less than or equal to 25), the number of home study supports (*hspprt*, with three levels; neither own room nor Internet connection, one of them *hspprt.1*, and both *hspprt.2*), student's confidence in science (*sci.conf*), and student's perceived value of science (*sci.value*). The school-level covariates were school's gender type (*gender.type*, with three levels; all-boys, all-girls *girl.sch*, and co-education *coedu*), the percentage of economically disadvantaged students (*pct.disad*, with four levels; 0 to 10%, 11 to 25% *disad.11*, 26 to 50% *disad.26*, and more than 50% *disad.M50*), school location (*city.size*, with four levels; urban *city.U*, suburban *city.Sub*, medium size city *city.M*, and small town), science instruction affected by resource shortage (*res.short*), school's emphasis on academic success (*aca.emph*), and school discipline problems (*dscpn*).

5.2 Results

In the first step of our hybrid ML methods, we determined the optimal number of latent classes by comparing the AIC measures under different numbers of latent classes. The two-class model had the lowest AIC and Latent Class 1 had 1,556 students from 44 schools, and Latent Class 2 had 3,318 students from 105 schools. Latent Class 1 was about 60% smaller than Latent Class 2; Latent Class 1 had 31.9% of the total students from 29.5% of the schools, and Latent Class 2 had 68.1% of the total students from 70.5% of the schools.

Figure 3 plots the estimated selection models $\widehat{e}_k$ from each latent class as a function of two observed covariates, value in science and resource shortages. The figure contains the line of best fit to guide visualization. For Latent Class 2, the propensity of taking private lessons was linearly increasing with how much value students placed in the sciences (*sci.value*). However, for Latent Class 1, the propensity remained flat and there was no discernable relationship between *sci.value* and selection probabilities. We also observed that a cluster-level covariate *res.short* increased the selection probabilities in Class 1, but there was no increasing pattern in Class 2.

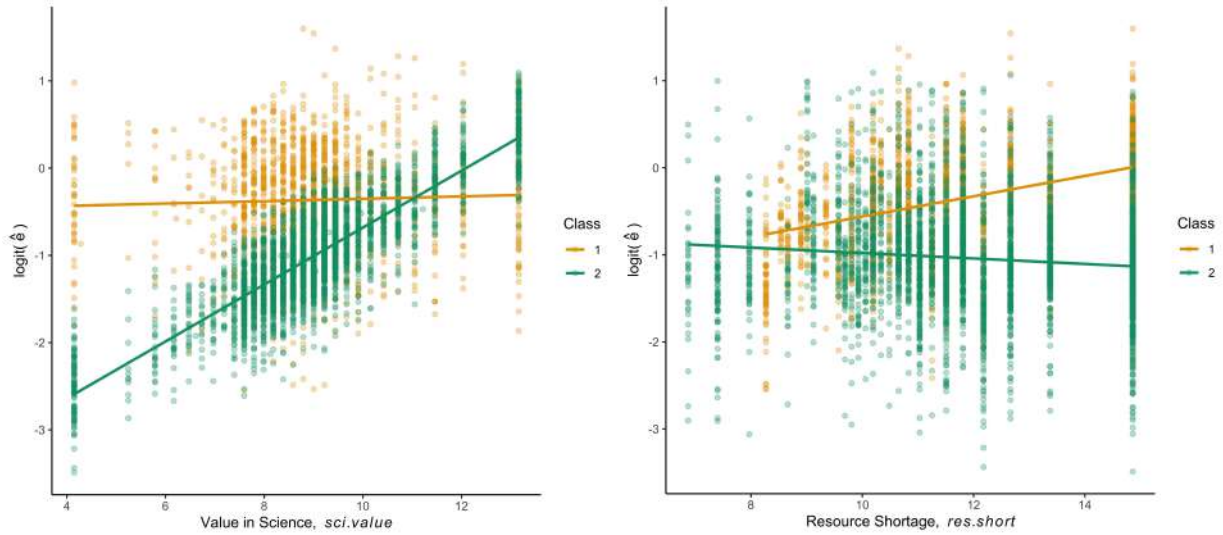

Figure 3: Class-specific selection models with respect to individual-level and cluster-level covariates. Each dot indicates a students' estimated logit propensity score, $logit(\hat{e}_k)$. Line of best fit is plotted to guide visualization.

We also summarize the student-level and school-level variables in Table 2. For most covariates, the two latent classes showed similar descriptive statistics. However, we found statistically significant differences in the propensity of taking private science lessons and father's educational level. Students in Class 1 were more likely to have a higher probability of taking private lessons and come from families whose father did not hold a college degree than those in Class 2. Looking at both Table 2 and Figure 3, students in Class 1 likely sought private lessons because they may received inadequate lessons at schools, whereas those in Class 2 likely sought private lessons because their families had stronger education backgrounds and may had placed a high value in science. Overall, as suspected from our subject-matter expertise, the latent

Table 2: Descriptive statistics of the two latent classes

	Class 1		Class 2	
	Mean or Percent	Std Dev	Mean or Percent	Std Dev
Student-level Variables	(N=1,556 students)		(N=3,318 students)	
science.score	557.84	75.94	555.97	76.25
math.score	608.24	82.74	605.13	84.63
propensity.score	**0.42**		**0.28**	
sci.conf	8.68	2.09	8.61	2.11
sci.value	8.98	1.64	8.92	1.64
male	51.6%		48.6%	
dad.cll	**34.4%**		**37.5%**	
dad.q	28.2%		28.1%	
books25	86.1%		86.0%	
hssprt.2	70.2%		71.9%	
School-level Variables	(J=44 schools)		(J=105 schools)	
res.short	11.71	2.04	11.79	2.00
aca.demph	11.14	1.84	11.10	1.87
dscpn	10.74	2.07	11.16	1.99
girl.sch	15.9%		21.9%	
coedu	65.9%		58.1%	
disad.11	29.5%		37.1%	
disad.26	27.3%		23.8%	
disad.M50	11.4%		10.5%	
city.U	40.9%		35.2%	
city.Sub	6.8%		9.5%	
city.M	31.8%		27.6%	

Note: Values in bold are significant differences between classes at α=0.05.

class selection model revealed different latent structures where students in each latent class had different propensities to seek private tutors.

Figure 4 shows the distributions of individual CATE estimates from Hybrid Causal Forests. The figure also shows vanilla Causal Forests that did not consider latent class membership. Using Hybrid Causal Forests, we see two different distributions centering around one and ten, respectively, to reflect variation in treatment effects between latent classes. In contrast, the vanilla Causal Forests only shows variation in treatment effects in the observed covariates and centers around four. Appendix E shows the distributions of CATE based on other ML methods.

Table 3 summarizes average treatment effect estimates of private science lessons within each latent class. As a comparison, we used within-class IPW estimator, MMW-S, and DR estimator with parametric propensity score or outcome models to estimate class-specific average treatment effects; we remark that hybrid ML methods do not require a priori specification of the propensity score or the outcome model. We included covariate balance plots in Appendix F. After estimating the average treatment effect within latent classes, we found that the prima facie effects amounted to 16.99 and 20.24 for Class 1 and Class 2, respectively. The prima facie effect is the unadjusted mean difference in science achievement scores between the treated and untreated groups. The treatment effects with IPW, MMW-S, and DR estimators varied depending on the latent class; there were significantly positive effects in Class 1, while no significant effects existed in Class 2. When we implemented hybrid ML methods, we observed that the average treatment effect estimates in both classes were similar to parametric methods except that the estimates for Class 2 were generally smaller than parametric methods.

For the one class model (far right column) where we assumed no latent class structure, the prima facie effect amounted to 18.96 points. After applying MMW-S, the effect decreased to 1.73 points, which was not statistically significant. However, one-class IPW and DR methods produced positive, significant, but reduced effects. Also, the average treatment effects with hybrid ML methods were positive, but slightly

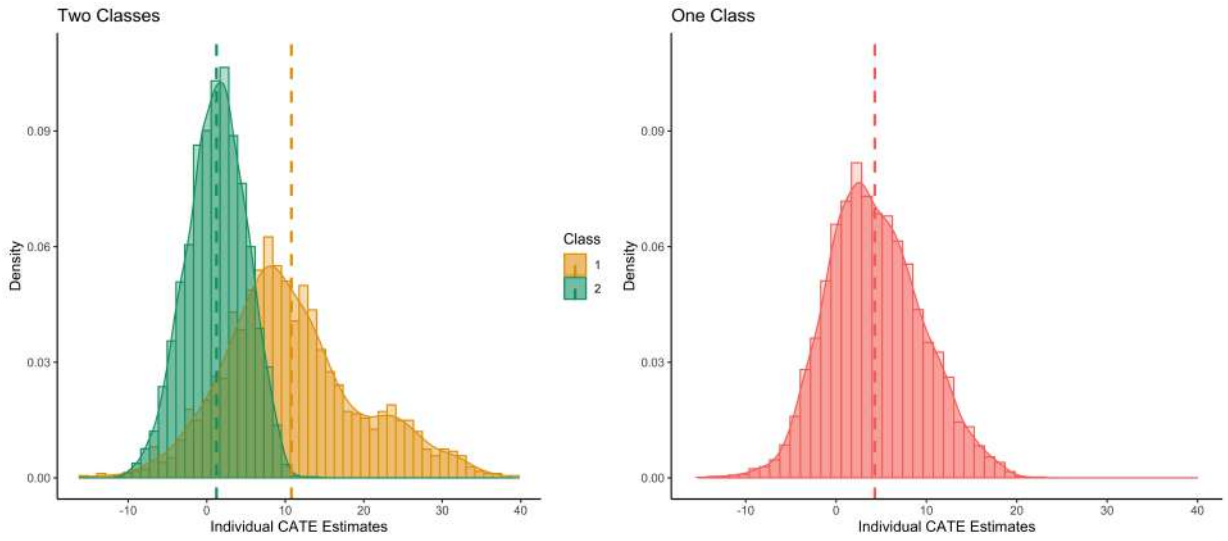

Figure 4: Distributions of individual CATE estimates from Causal Forests. The left shows Hybrid Causal Forests discovering two latent classes, while the right shows vanilla Causal Forests without consideration for latent classes. Dashed lines represent class-specific treatment effect estimates

Table 3: Comparisons of the class-specific average treatment effect estimates

	Two Classes				One Class	
	Class 1		Class 2			
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Prima facie (unadjusted)	**16.99**	(3.88)	**20.24**	(2.93)	**18.96**	(2.32)
IPW	**11.62**	(2.58)	2.93	(2.94)	**5.36**	(2.06)
MMW-S	**12.28**	(2.69)	2.33	(2.78)	1.73	(2.08)
DR	**11.78**	(2.64)	3.02	(2.89)	**5.39**	(1.99)
Hybrid Causal Forests	**10.79**	(2.11)	1.24	(2.14)	**4.28**	(1.58)
Hybrid BART	**12.24**	(3.26)	0.73	(2.43)	**4.36**	(1.90)
Hybrid TMLE	**12.19**	(3.13)	1.40	(2.39)	**4.54**	(1.87)

Note: Standard errors (SE) were estimated using bootstrap sampling with 5,000 repetitions. Estimates in bold are significant at α=0.05. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents the doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation.

smaller than IPW and DR estimators. However, as mentioned before, none of these effects uncovered the potential effect heterogeneity within latent classes.

6 Discussion and Conclusions

We propose hybrid ML methods to estimate heterogeneous treatment effects between latent classes. Our proposed hybrid approach uses context-specific finite mixture models to identify different latent classes and ML-based causal inference methods to estimate treatment effects within each class. Broadly speaking, hybrid ML methods extend the capacities of ML methods to capture treatment effect heterogeneity defined by latent class mixture models. Our simulation study revealed that hybrid ML methods are an attractive alternative to existing propensity score methods. Finally, in our data analysis, we demonstrated that hybrid ML methods were able to capture heterogeneous effects and the average treatment effect for each latent class.

We make three concluding remarks about hybrid ML methods. First, ensuring sufficient sample sizes is important when using multilevel latent class mixture models. We observed that increasing cluster sizes affected the proportions of correctly identifying class membership and we generally recommend using hybrid ML methods when the number of clusters is more than 50 and the mean cluster size is more

than 30 so that the total sample size is at least 1500, the minimum sample size in our simulation design. Hybrid ML methods did perform well in our simulation study even when the maximum misclassification rate was about 27%. However, in general, if there are insufficient samples, there is an increased likelihood of misclassifying units and consequently, an increased risk of biasing the average treatment effect. Second, though ML methods can flexibly fit the outcome model and the propensity score model, this does not give a free pass for mis-classification in latent class models, and it would be an interesting topic of future research to design ML methods to be robust to biases arising from mis-classification in latent class models. Third, we believe that our work here provides a more systematic approach of applying ML methods for causal inference to education and psychology. In particular, we hope that the work provides a template for researchers to combine other types of latent class modeling with any ML-based causal inference methods to better understand the nature of treatment effect heterogeneity and the underlying latent structures in the data.

Notes

1. Academic resilience is defined as "the heightened likelihood of success in school and in other life accomplishments, despite environmental adversities brought about by early traits, conditions, and experiences" (Wang & Gordon, 2012).
2. Self-efficacy is defined as "people's beliefs in their ability to influence events that affect their lives" (Bandura, 2010).

References

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

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology*, *43*(1), 272–311.

Bandura, A. (2010). Self-efficacy. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology* (4th ed., pp. 1–3). Hoboken, NJ : Wiley.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Butera, N. M., Lanza, S. T., & Coffman, D. L. (2014). A framework for estimating causal effects in latent class analysis: Is there a causal link between early sex and subsequent profiles of delinquency? *Prevention science*, *15*(3), 397–407.

Clogg, C. C. (1995). Latent class models. In S. M. E. Arminger G. Clogg C. C. (Ed.), *Handbook of statistical modeling for the social and behavioral sciences* (p. 311-359). Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dorie, V., & Hill, J. (2019). bartcause: Causal inference using bayesian additive regression trees [Computer software manual]. Retrieved from `https://github.com/vdorie/bartCause` (R package version 1.0-0)

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

Foy, P., Arora, A., & Stanco, G. (2017). *Timss 2015 user guide for the international database*. TIMSS & PIRLS International Study Center, Boston College.

Goodman, L. A. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, *37*(1), 1–22.

Gruber, S., & van der Laan, M. J. (2012). tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, *51*(13), 1–35. Retrieved from `http://www.jstatsoft.org/v51/i13/` (doi:10.18637/jss.v051.i13)

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

Hallquist, M., & Wiley, J. (2017). Mplusautomation: Automating mplus model estimation and interpretation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=MplusAutomation` (R package version 0.7)

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis*, *31*(1), 54–81.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jo, B., Wang, C.-P., & Ialongo, N. S. (2009). Using latent outcome trajectory classes in causal inference. *Statistics and its Interface*, *2*(4), 403.

Kang, J., & Schafer, J. L. (2010). *Estimating average treatment effects when the treatment is a latent class* (Tech. Rep. No. 1005). Department of Statistics, The Pennsylvania State University.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (pp. 592–612). Sage.

Kim, J.-S., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In L. van der Ark, D. Bolt, W.-C. Wang, J. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research: The 80th annual meeting of the psychometric society* (pp. 293–306). Springer.

Kim, J.-S., Steiner, P. M., & Lim, W. C. (2016). Mixture modeling methods for causal inference with multilevel data. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 335–359). Information Age Publishing.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lanza, S. T., Coffman, D. L., & Xu, S. (2013). Causal inference in latent class analysis. *Structural equation modeling: a multidisciplinary journal*, *20*(3), 361–383.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent glass regression in r. *Journal of Statistical Software*, *11*.

Leite, W. (2016). *Practical propensity score methods using R*. Sage Publications.

Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, *65*(1), 93–119.

Ma, X. (2018). *Using classification and regression trees: A practical primer*. Information Age Publishing.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 175–198).

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in timss 2015*. TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Masyn, K. E. (2013). latent class analysis and finite mixture modeling. In T. Little (Ed.), *The oxford handbook of quantitative methods* (p. 551-611). New York, NY: Oxford University Press.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.

Mok, J.-Y., Choi, S.-W., Kim, D.-J., Choi, J.-S., Lee, J., Ahn, H., ... Song, W.-Y. (2014). Latent class analysis on internet and smartphone addiction in college students. *Neuropsychiatric disease and treatment*, *10*, 817.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments: essay on principles. section 9 (with discussion). *Statistical Science*, *4*, 465–480.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961–962.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, *13*(4), 279.

Schuler, M. S., Leoutsakos, J.-M. S., & Stuart, E. A. (2014). Addressing confounding when estimating the effects of latent classes on a distal outcome. *Health Services and Outcomes Research Methodology*, *14*(4), 232–254.

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. Little (Ed.), *The oxford handbook of quantitative methods* (p. 236-258). New York, NY: Oxford University Press.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*(5), 141–158.

Suk, Y., Kang, H., & Kim, J.-S. (2019, Sep). *Random forests approach for causal inference with clustered observational data*. PsyArXiv. Retrieved from psyarxiv.com/xgq2k doi: 10.31234/osf.io/xgq2k

Sutfin, E. L., Reboussin, B. A., McCoy, T. P., & Wolfson, M. (2009). Are college student smokers really a homogeneous group? a latent class analysis of college student smokers. *Nicotine & Tobacco Research*, *11*(4), 444–454.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Miner, L., Wager, S., & Wright, M. (2019). grf: Generalized random forests (beta) [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=grf (R package version 0.10.3)

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley.

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling*, *18*(1), 110–131.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1).

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological methodology*, *33*(1), 213–239.

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

Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis*, *41*(3-4), 531–537.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using mplus*. John Wiley & Sons.

Wang, M. C., & Gordon, E. W. (2012). *Educational resilience in inner-city america: Challenges and prospects*. Routledge.

A Comparison between the sequential approach and the covariate approach

We compared the performance between our proposed sequential/two-step approach and an alternative "covariate" approach where the estimated class membership is used as an additional covariate. Specifically, the proposed approach implements ML methods within each latent class and estimates the ATE within each class, while the covariate approach includes the estimated class membership variable as another covariate in ML methods and estimates the conditional ATE defined by this covariate. We suspect that under some assumptions, both are asymptotically equivalent, but they may have different finite-sample properties and we investigate them through a small simulation study below.

Table 4 provides the performance of class-specific ATE estimates between the the two approaches using the data generating model from the main text. We saw that biases from our two-step approach were generally smaller than those from the covariate approach. But as the cluster size increased, biases became smaller for both approaches and the differences between the two were generally negligible. Also, the MSEs were consistently smaller under our approach across different cluster sizes. While the simulation study is small, the result gives some confidence that our approach outperforms the covariate approach in terms of finite-sample bias and MSE.

Table 4: Comparison between the proposed sequential approach and the covariate approach

	Latent Class 1		Latent Class 2	
(nC1, nC2, nS)	|Bias|	MSE	|Bias|	MSE
Our Approach				
(25, 25, 50)	0.141	0.538	0.088	0.403
(25, 25, 100)	0.039	0.259	0.073	0.186
(25, 25, 200)	0.001	0.111	0.018	0.101
(25, 25, 400)	0.014	0.062	0.013	0.051
Covariate Approach				
(25, 25, 50)	0.669	0.844	0.711	0.828
(25, 25, 100)	0.562	0.560	0.525	0.470
(25, 25, 200)	0.309	0.247	0.210	0.179
(25, 25, 400)	0.111	0.092	0.075	0.067

B Performance of class-specific treatment effect estimates with estimated class membership

Table 5: Performance of class-specific treatment effect estimates: estimated class membership

	Latent Class 1		Latent Class 2		Overall	
(nC1, nC2, nS)	|Bias|	MSE	|Bias|	MSE	|Bias|	MSE
(25, 25, 30)						
IPW	0.009	1.825	0.802	1.762	0.405	1.790
MMW-S	0.228	1.920	0.788	1.840	0.280	1.876
DR	0.271	1.393	0.832	1.896	0.280	1.640
Hybrid Causal Forests	0.373	1.302	0.762	1.746	0.194	1.521
Hybrid BART	0.482	1.274	0.751	1.660	0.134	1.464
Hybrid TMLE	0.272	1.150	0.757	1.756	0.242	1.449
(25, 25, 50)						
IPW	0.004	1.039	0.469	0.944	0.236	0.988
MMW-S	0.186	0.914	0.441	0.849	0.127	0.879
DR	0.174	0.803	0.472	0.967	0.148	0.882
Hybrid Causal Forests	0.233	0.719	0.399	0.739	0.082	0.727
Hybrid BART	0.297	0.660	0.411	0.762	0.057	0.710
Hybrid TMLE	0.179	0.638	0.427	0.802	0.123	0.718
(50, 50, 30)						
IPW	0.043	0.818	0.559	0.797	0.258	0.806
MMW-S	0.416	1.074	0.541	0.790	0.063	0.931
DR	0.343	0.645	0.594	1.014	0.126	0.830
Hybrid Causal Forests	0.387	0.696	0.533	0.762	0.073	0.728
Hybrid BART	0.455	0.689	0.519	0.697	0.032	0.692
Hybrid TMLE	0.337	0.592	0.525	0.743	0.095	0.667

Note: nC1, nC2, and nS represent the number of clusters for the first latent class, the number of clusters for the second latent class, and average cluster sizes, respectively. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents a doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation. The true treatment effect values are 2.5 and 0 for the first and second latent classes, respectively.

C Performance of class-specific treatment effect estimates with true class membership

We replicated the simulation study from the main text and measured the performance of class-specific treatment effect estimates when we use the true class labels. In this setting, all the parametric methods have correctly specified outcome and propensity score models and should perform well. Specifically, we expect the DR estimator to perform the best followed by the non-DR estimators (IPW and MMW-S). Finally, we expect the performance of ML methods to be somewhere in between the performance of the DR and non-DR estimators, but the performance of ML methods will become similar to the performance of the DR estimator as the sample size increases.

Table 6 shows the results. As expected, the DR estimators performed best in terms of bias and MSE. The non-DR estimators—IPW and MMW-S—performed worse than the DR estimator and ML methods with one exception: absolute bias of the MMW-S in the sample size condition (25, 25, 50). Also, similar to what we observed in the main text, we saw that the non-DR estimators, especially the IPW estimator, achieved more bias reduction in one latent class over another latent class in the current data generating model, but ended up having a relatively large amount of overall bias. In contrast, DR and hybrid methods achieved bias reduction in both latent classes and had overall bias reductions. Finally, when the sample size increased, the performance of ML methods was competitive to the performance of the DR estimator.

Table 6: Performance of class-specific treatment effect estimates: true class membership

	Latent Class 1		Latent Class 2		Overall	
(nC1, nC2, nS)	|Bias|	MSE	|Bias|	MSE	|Bias|	MSE
(25, 25, 30)						
IPW	0.371	1.521	0.008	0.586	0.182	1.054
MMW-S	0.138	1.455	0.064	0.625	0.101	1.040
DR	0.032	0.732	0.018	0.595	0.007	0.663
Hybrid Causal Forests	0.065	0.848	0.010	0.634	0.028	0.741
Hybrid BART	0.048	0.736	0.023	0.595	0.036	0.665
Hybrid TMLE	0.114	0.741	0.007	0.609	0.054	0.675
(25, 25, 50)						
IPW	0.227	0.964	0.003	0.342	0.115	0.652
MMW-S	0.005	0.787	0.025	0.349	0.016	0.568
DR	0.037	0.601	0.000	0.344	0.018	0.473
Hybrid Causal Forests	0.063	0.553	0.003	0.353	0.030	0.453
Hybrid BART	0.001	0.524	0.008	0.343	0.005	0.434
Hybrid TMLE	0.086	0.533	0.007	0.347	0.046	0.440
(50, 50, 30)						
IPW	0.313	0.747	0.025	0.281	0.144	0.514
MMW-S	0.222	0.737	0.083	0.297	0.153	0.517
DR	0.019	0.374	0.034	0.269	0.027	0.322
Hybrid Causal Forests	0.072	0.417	0.038	0.289	0.017	0.353
Hybrid BART	0.034	0.384	0.042	0.282	0.038	0.333
Hybrid TMLE	0.046	0.362	0.034	0.279	0.006	0.320

Note: nC1, nC2, and nS represent the number of clusters for the first latent class, the number of clusters for the second latent class, and average cluster sizes, respectively. IPW represents inverse-propensity weighting, and MMW-S represents marginal mean weighting through stratification. DR represents a doubly robust estimator. BART represents Bayesian additive regression trees, and TMLE represents targeted maximum likelihood estimation. The true treatment effect values are 2.5 and 0 for the first and second latent classes, respectively.

D Evaluation of individual CATE estimates

We also assessed the performance of ML methods based on the root mean squared error (RMSE) of estimating individual CATE estimates in simulation replication m, denoted as $\hat{\tau}_{ij,m}(k)$. Specifically, let N_m denote the sample size of each simulation replication. We evaluated the following quantity:

$$\text{RMSE}_m(k) = \sqrt{\frac{1}{N_m}\sum_{ij}(\hat{\tau}_{ij,m}(k) - \tau_{ij,m}(k))^2}$$

and took averages of $\text{RMSE}_m(k)$ across simulation replicates.

Figure 5 summarizes the performance of individual CATE estimates within each latent class across different ML methods. Though TMLE tends to have slightly large RMSEs, the performance across methods was comparable and our simulation results in the main manuscript are generally not sensitive to the choice of ML methods.

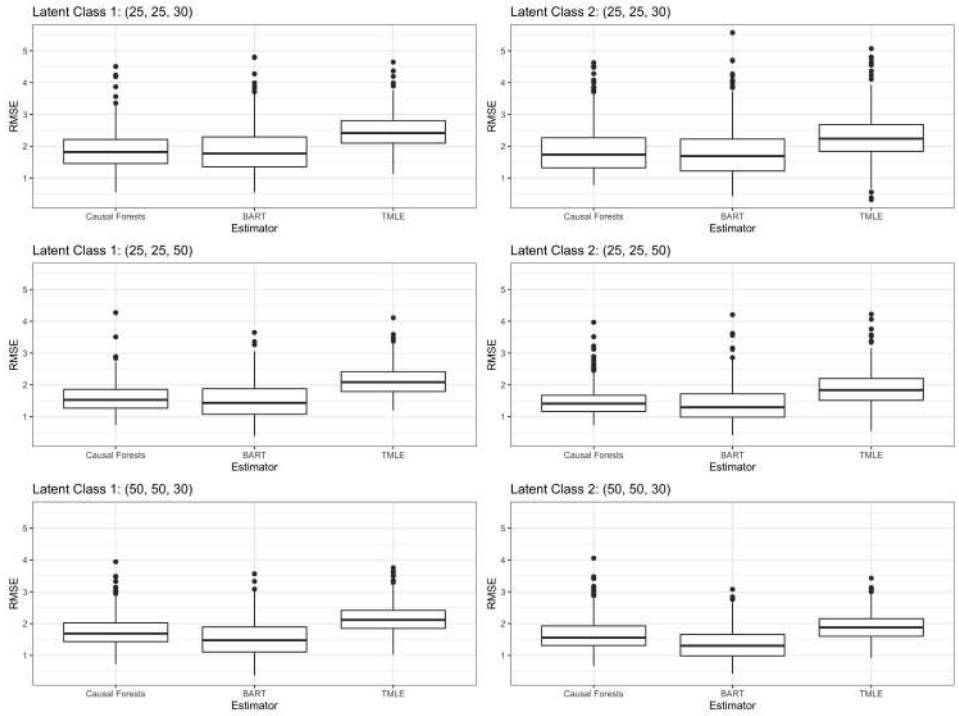

Figure 5: Performance of individual CATE estimates: root mean squared error.

E Distributions of individual CATE estimates

Figure 6 displays the distributions of individual CATE estimates when BART and TMLE are used. Our hybrid approach with different ML methods produced similar estimates of $\tau(k)$; the dotted lines across methods align closely with each other. However, there were some differences in the distributions of the individual CATE estimates, with BART producing "spiky" distributions, whereas TMLE producing smoother distributions. This suggests that different ML methods make different assumptions about how to locally smooth across the observed covariates.

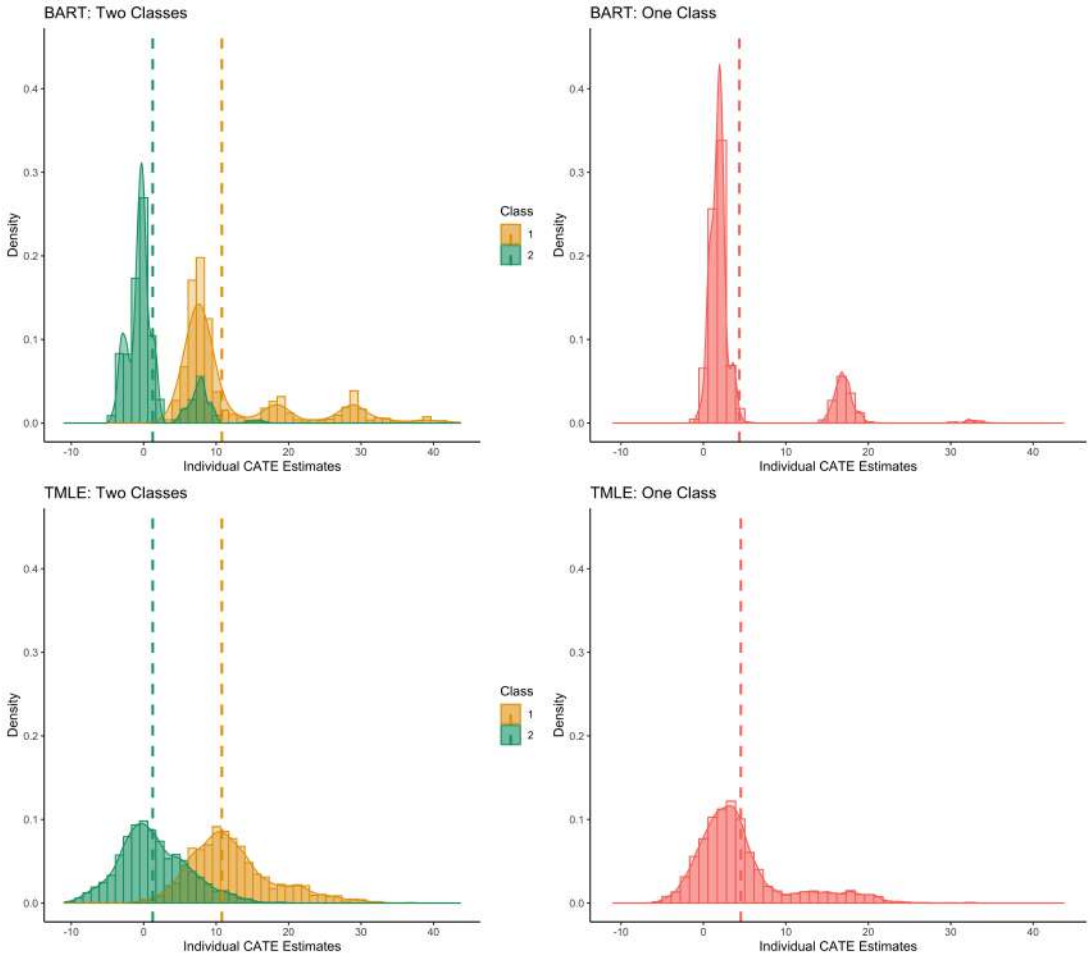

Figure 6: Distributions of individual CATE estimates with BART and TMLE. The left shows hybrid ML methods discovering two latent classes, while the right shows the usual ML methods without latent classes. Dashed lines represent class-specific treatment effect estimates

F Covariance Balance

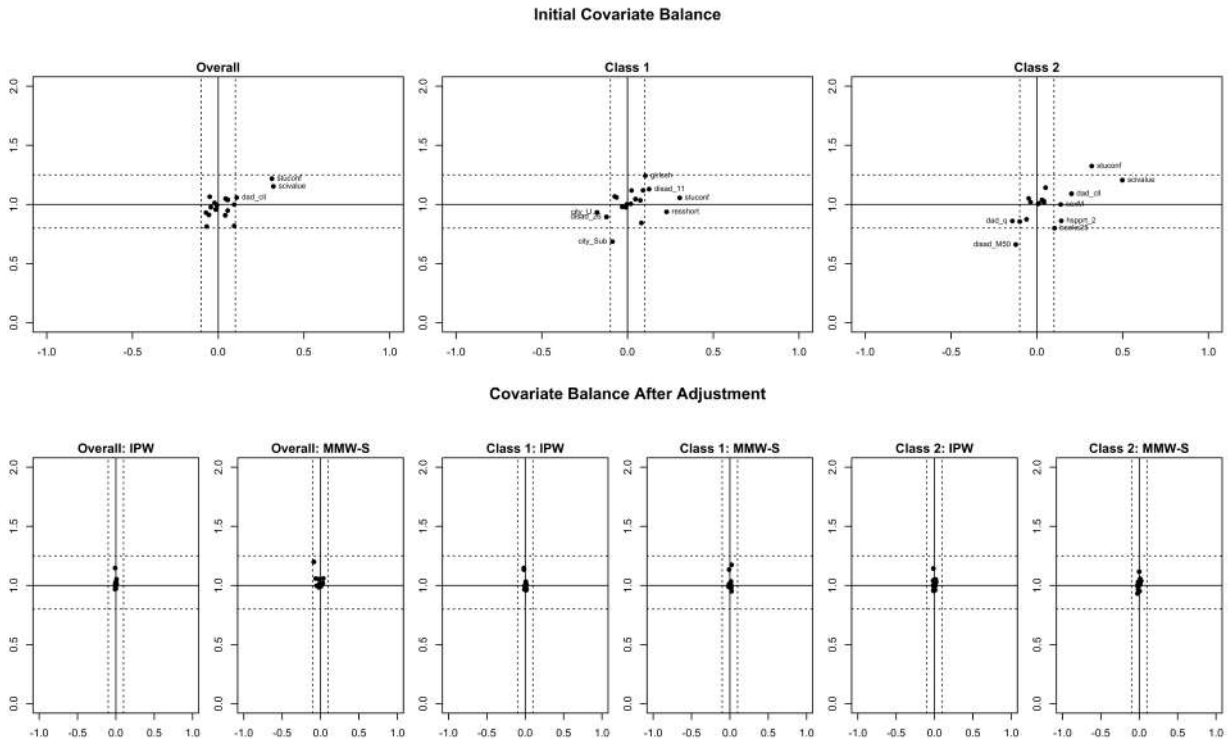

Figure 7: Covariate balance plots before and after propensity score adjustment
(Standardized mean differences on the x-axis and variance ratios on the y-axis)

We checked covariate balance in within-class matching by computing the absolute standardized mean differences and variance ratios between treated units and control units. As a rule of thumb, if the mean difference of each covariate is less than 0.1 standard deviation and the variance ratio is more than 4/5 and less than 5/4, we can provide evidence for good balance of the covariates. Figure 7 displays covariate balance plots before and after propensity score adjustment for two classes as well as for one homogeneous class. One homogeneous class assumed no subpopulations or multiple latent classes, and its propensity scores were estimated via random effects logistic regression. There was less initial imbalance in covariates for one homogeneous class, and we achieved good covariate balance between the treated and untreated groups after applying IPW and MMW-S. For the two-class approach, the covariates imbalanced differed in each class. However, after applying IPW and MMW-S, we achieved acceptable covariate balance within each class.