Estimation of Local Average Treatment Effect by Data Combination

Kazuhiko Shinoda[1], Takahiro Hoshino[2,3]

[1]Graduate School of Economics, Keio University
[2]Faculty of Economics, Keio University
[3]RIKEN AIP
k-shinoda@keio.jp, bayesian@jasmine.ocn.ne.jp

Abstract

It is important to estimate the local average treatment effect (LATE) when compliance with a treatment assignment is incomplete. The previously proposed methods for LATE estimation required all relevant variables to be jointly observed in a single dataset; however, it is sometimes difficult or even impossible to collect such data in many real-world problems for technical or privacy reasons. We consider a novel problem setting in which LATE, as a function of covariates, is nonparametrically identified from the combination of separately observed datasets. For estimation, we show that the direct least squares method, which was originally developed for estimating the average treatment effect under complete compliance, is applicable to our setting. However, model selection and hyperparameter tuning for the direct least squares estimator can be unstable in practice since it is defined as a solution to the minimax problem. We then propose a weighted least squares estimator that enables simpler model selection by avoiding the minimax objective formulation. Unlike the inverse probability weighted (IPW) estimator, the proposed estimator directly uses the pre-estimated weight without inversion, avoiding the problems caused by the IPW methods. We demonstrate the effectiveness of our method through experiments using synthetic and real-world datasets.

1 Introduction

Estimating the causal effects of treatment on an outcome of interest is central to optimal decision making in many real-world problems, such as policymaking, epidemiology, education, and marketing (Skovron and Titiunik 2015; Wood et al. 2008; Oreopoulos 2006; Varian 2016). However, the identification and estimation of treatment effects usually relies on the untestable assumption referred to as *unconfoundedness*, namely, independence between the treatment status and potential outcomes (Imbens and Rubin 2015). Violations of unconfoundedness may occur not only in observational studies, but also in randomized controlled trials (RCTs) when compliance to the assigned treatment is not complete. For example, even if a coupon is distributed randomly to measure its effect on sales, the probability of using the coupon is likely to depend on the unobserved nature of the individuals. Moreover, noncompliance can also occur

regardless of the individual's intentions. In online advertisement placement, the probability of watching the ad depends on the bidding strategy of other companies because ads that are actually displayed are determined through the real-time-bidding even if one tries to randomly place the ad.

In such cases, it is well known that *the local average treatment effect (LATE)* can be identified and estimated using the treatment assignment as an instrumental variable (IV) and conditions milder than unconfoundedness (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996; Frölich 2007). LATE is the treatment effect measured for the subpopulation of compliers, individuals who always follow the given assignment.

We suppose that we cannot observe all relevant variables in a single dataset for technical or privacy reasons. For example, in online-to-offline (O2O) marketing, where treatments are implemented online and outcomes are observed offline, it is often difficult to match the records of the same individuals observed separately online and offline. Additionally, with the global anti-tracking movement gaining momentum, it may become more difficult to combine multiple pieces of information online as well. Although causal inference by data combination has been actively studied (Ridder and Moffitt 2007; Bareinboim and Pearl 2016; Lee, Correa, and Bareinboim 2020), LATE estimation using multiple datasets has not received much attention despite its practical importance. We extend the problem setting considered in (Yamane et al. 2018), where two different treatment regimes are available, to allow for the existence of noncompliance.

For the estimation, we show that the direct estimation method originally developed for the average treatment effect (ATE) under the complete compliance (Yamane et al. 2018) can be applied to the LATE estimation in our setting. However, their method has a practical issue in that model selection and hyperparameter tuning can be unstable owing to its minimax objective formulation. We then propose a weighted least squares estimator to avoid the minimax objective, and improve the stability in practice. Unlike the inverse probability weighted (IPW) estimator (Wooldridge 2002, 2007; Seaman and White 2013), which is often employed to estimate treatment effects, the proposed estimator directly uses the estimated propensity-score-difference (PSD) as a weight without inversion. Therefore, our method can also avoid the common issue in the IPW methods, that is, high variance at

points with a propensity score extremely close to zero.

The contributions of this study lie in the following three parts. First, we show that LATE is identified even when an outcome and treatment status cannot be observed simultaneously in a single dataset, and the treatment assignment is completely missing. Second, we find that the positivity assumption, which is necessary in the standard setting with one regime, can be omitted in our setting. We show this relaxation of the conditions further facilitates data collection. Third, we develop a novel estimator that enables simpler model selection while maintaining the essence of direct estimation as much as possible.

2 Problem Setting

Unlike the standard causal inference studies, we consider a setting where there are two different assignment regimes (Yamane et al. 2018), which we term as *the two-regime design (2RD)*. The concept is quiet general that it only requires two observational studies, two RCTs or a combination of the two. We have to be assured that the treatment assignment is done based on different regimes, i.e. different probabilities (see Assumption 1.2).

We define our problem using the potential outcome framework (Rubin 1974; Imbens and Rubin 2015). Let $K \in \{0, 1\}$ be a regime indicator, and we use a superscript $k = 0, 1$ to specify the regime which the variables come from. Let $Y^{(k)} \in \mathcal{Y} \subset \mathbb{R}$ be an outcome of interest, $D^{(k)} \in \{0, 1\}$ be a binary treatment indicator, $Z^{(k)} \in \{0, 1\}$ be an assignment indicator and $\boldsymbol{X}^{(k)} \in \mathcal{X} \subset \mathbb{R}^{q_x}$ be a q_x-dimensional vector of covariates. $D_z^{(k)}$ is the potential treatment status realized only when $Z^{(k)} = z$, and $Y_d^{(k)}$ is the potential outcome realized only when $D^{(k)} = d$, where $z, d \in \{0, 1\}$. Using this notation, we implicitly assume that $Z^{(k)}$ does not have a direct effect on $Y^{(k)}$, but affects $Y^{(k)}$ indirectly through $D^{(k)}$. This condition, often referred to as the exclusion restriction, is necessary for $Z^{(k)}$ to be a valid IV. Let Y_1, Y_0, D_1, D_0 and $\boldsymbol{X}$ be the potential variables and covariates in the population of interest, and we suppose that the iid samples from $P(\boldsymbol{X})$ can be obtained as test samples.

2.1 Basic Setting

First, we make the following assumption on the relation between the two regimes.

Assumption 1.

1. $P(Y_1^{(1)}, Y_0^{(1)}, D_1^{(1)}, D_0^{(1)}, \boldsymbol{X}^{(1)})$
   $= P(Y_1^{(0)}, Y_0^{(0)}, D_1^{(0)}, D_0^{(0)}, \boldsymbol{X}^{(0)})$
   $= P(Y_1, Y_0, D_1, D_0, \boldsymbol{X})$.
2. $P(Z^{(1)} = 1|\boldsymbol{X}^{(1)} = \boldsymbol{x}) \neq P(Z^{(0)} = 1|\boldsymbol{X}^{(0)} = \boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$.

By Assumption 1.1, we suppose that the joint distribution of the potential variables and covariates in each assignment regime is invariant to each other, and equal to the joint distribution in the population of interest. This means that the participants in each regime are random draws from the population of interest. We can still identify LATE even if we weaken Assumption 1.1 to its conditional version, but the

$$P(Y_1, Y_0, D_1, D_0, \boldsymbol{X})$$
$$K = 1 \qquad\qquad K = 0$$
Regime Assignment
$$P(Y_1^{(1)}, Y_0^{(1)}, D_1^{(1)}, D_0^{(1)}, \boldsymbol{X}^{(1)}) \qquad P(Y_1^{(0)}, Y_0^{(0)}, D_1^{(0)}, D_0^{(0)}, \boldsymbol{X}^{(0)})$$
Treatment Assignment with
$$P(Z^{(1)} = 1|\boldsymbol{X}^{(1)}) \neq P(Z^{(0)} = 1|\boldsymbol{X}^{(0)})$$
$$P(Y^{(1)}, D^{(1)}, \boldsymbol{X}^{(1)}) \qquad\qquad P(Y^{(0)}, D^{(0)}, \boldsymbol{X}^{(0)})$$
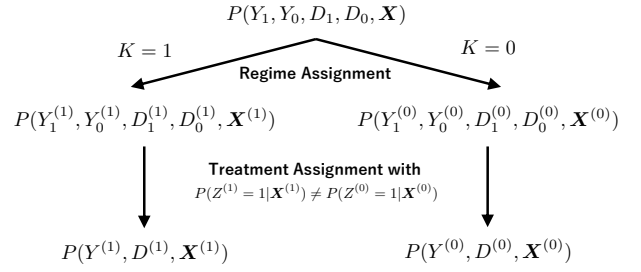
Figure 1: Overview of the 2RD.

direct estimation is no longer possible. See Appendix B in the supplementary material for more discussion. Assumption 1.2 indicates that the assignment regimes are different to each other for any level of the covariates.

Additionally, we make the following assumption necessary for the identification of LATE. Here, $A \perp\!\!\!\perp B|C$ means A and B are conditionally independent given C.

Assumption 2. For $k = 0, 1$,

1. $Y_1^{(k)}, Y_0^{(k)}, D_1^{(k)}, D_0^{(k)} \perp\!\!\!\perp Z^{(k)}|\boldsymbol{X}^{(k)}$.
2. $D^{(k)} = Z^{(k)}D_1^{(k)} + (1 - Z^{(k)})D_0^{(k)}$ and $Y^{(k)} = D^{(k)}Y_1^{(k)} + (1 - D^{(k)})Y_0^{(k)}$.
3. $P(D_1^{(k)} \geq D_0^{(k)}|\boldsymbol{X}^{(k)}) = 1$.
4. $P(D_1^{(k)} = 1|\boldsymbol{X}^{(k)}) \neq P(D_0^{(k)} = 1|\boldsymbol{X}^{(k)})$.

Assumption 2.1 states that $Z^{(k)}$ are randomly assigned within a subpopulation sharing the same level of the covariates. The mean independence between the potential variables and the treatment assignment conditional on covariates is sufficient for identifying LATE, but we maintain Assumption 2.1 as it is typical to assume the full conditional independence in the causal inference literature. Assumption 2.2 relates the potential variables to their realized counterparts. Assumption 2.3 excludes defiers, those who never follow the given treatment assignment, from our analysis. Assumption 2.4 is necessary for $Z^{(k)}$ to be a valid IV of $D^{(k)}$. Figure 1 illustrates the data generating process under the 2RD.

Assumption 2 is the direct extension of the standard assumptions used for the LATE estimation (Abadie 2003; Frölich 2007; Tan 2006; Ogburn, Rotnitzky, and Robins 2015) to the 2RD. However, we omit the condition referred to as positivity $0 < P(Z^{(k)} = 1|\boldsymbol{X}^{(k)}) < 1$ as it is unnecessary in the 2RD. Therefore, it is possible to set $P(Z^{(0)} = 1|\boldsymbol{X}^{(0)}) = 0$ as long as Assumption 1.2 is satisfied. We term the design with $P(Z^{(1)} = 1|\boldsymbol{X}^{(1)}) > 0$ for any $\boldsymbol{X}^{(1)}$ and $P(Z^{(0)} = 1|\boldsymbol{X}^{(0)}) = 0$ for any $\boldsymbol{X}^{(0)}$ as *the one-experiment 2RD (1E2RD)* since an experiment is conducted only for those with $k = 1$, and a natural state without any intervention is observed for those with $k = 0$.

Although (Yamane et al. 2018) does not mention anything on this point, it is of great practical importance since the 1E2RD is much easier to implement than the general 2RD, and extends the plausibility of our setting. In the 1E2RD, we only have to conduct one experiment or collect one set of observational data just like the standard setting for causal

inference. All we need in addition is a dataset collected from those without any intervention. Such data may be available at no additional cost, for example where there is appropriate open data published by the government.

Hereafter, we abuse the notation by omitting the superscript on the potential variables and covariates unless this causes confusion.

2.2 Identification

The parameter of our interest is LATE as a function of covariates $\boldsymbol{X}$, defined as

$$\mu(\boldsymbol{X}) := E[Y_1 - Y_0 | D_1 > D_0, \boldsymbol{X}].$$

It measures how covariates $\boldsymbol{X}$ affect the average treatment effect within the subpopulation of compliers. The following theorem shows that $\mu(\boldsymbol{X})$ is nonparametrically identified in an unusual form in our setting.

Theorem 1. *Under Assumption 1 and 2,*

$$\mu(\boldsymbol{X}) = \frac{E[Y^{(1)} - Y^{(0)} | \boldsymbol{X}]}{E[D^{(1)} - D^{(0)} | \boldsymbol{X}]}. \tag{1}$$

The proof can be found in Appendix A in the supplementary material. Notably, this form coincides with the identification result of ATE in (Yamane et al. 2018), which means that we can also estimate $\mu(\boldsymbol{X})$ from the same combination of datasets proposed in (Yamane et al. 2018) under appropriate assumptions. This is a rather powerful result in practice since $Z^{(k)}$ is not required despite the presence of noncompliance. Moreover, it is obvious that $p_d^{(k)} := P(D^{(k)} = 1)$ and $P(\boldsymbol{X} | D^{(k)} = 1)$ are sufficient to identify the denominator of (1) since $E[D^{(k)} | \boldsymbol{X}]$ can be identified as $\frac{P(\boldsymbol{X} | D^{(k)}=1) p_d^{(k)}}{P(\boldsymbol{X})}$ by applying the Bayes' theorem.

The point is that we can identify $\mu(\boldsymbol{X})$ as a combination of functions depending only on covariates $\boldsymbol{X}$ in our setting. This property allows us to use a direct estimation method for $\mu(\boldsymbol{X})$. We can identify LATE as $\mu(\boldsymbol{X}) = \frac{E[Y|Z=1,\boldsymbol{X}] - E[Y|Z=0,\boldsymbol{X}]}{E[D|Z=1,\boldsymbol{X}] - E[D|Z=0,\boldsymbol{X}]}$ under the standard assumptions (Abadie 2003; Frölich 2007; Tan 2006; Ogburn, Rotnitzky, and Robins 2015), but we cannot rewrite it as a combination of functions depending only on covariates (see Appendix B).

Theorem 1 also suggests the usefulness of the 1E2RD. We can simplify (1) when we implement the 1E2RD under *one-sided noncompliance*, where individuals assigned to the control group never receive treatment, but they have a choice if assigned to the treatment group.

Corollary 1. *Assume $P(Z^{(0)} = 1 | \boldsymbol{X}) = 0$ (1E2RD) and $P(D_0 = 1 | \boldsymbol{X}) = 0$ (one-sided noncompliance) in addition to Assumption 1 and 2. Then, $E[Y^{(0)} | \boldsymbol{X}] = E[Y_0 | \boldsymbol{X}]$ and*

$$\mu(\boldsymbol{X}) = \frac{E[Y^{(1)} - Y^{(0)} | \boldsymbol{X}]}{E[D^{(1)} | \boldsymbol{X}]}.$$

This corollary shows that we can reduce the number of necessary datasets in the 1E2RD under one-sided noncompliance. This fact not only facilitates the data collection, but also benefits the estimation since the denominator is now just a propensity score, thus estimable accurately by,

for example, the Positive-Unlabeled (PU) learning (Elkan and Noto 2008; du Plessis, Niu, and Sugiyama 2014, 2015; Kiryo et al. 2017) with the logistic loss.

Although one-sided noncompliance is often associated with RCTs, some observational studies also fit with the framework (Frölich and Melly 2013; Kennedy 2020). In the case of one-sided noncompliance, we can also consider our problem as the estimation of the average treatment effect on the treated (ATT) since LATE is equal to ATT under one-sided noncompliance and the other standard assumptions (Frölich and Melly 2013; Donald, Hsu, and Lieli 2014).

2.3 Data Collection Scheme

We assume that the joint samples of $(Y^{(k)}, D^{(k)}, Z^{(k)}, \boldsymbol{X})$ are not available. By Theorem 1, the following separate datasets and the estimate of $p_d^{(k)}$ are sufficient for estimating $\mu(\boldsymbol{X})$:

$$\{\boldsymbol{x}_{di}^{(k)}\}_{i=1}^{n_d^{(k)}} \overset{\text{iid}}{\sim} P(\boldsymbol{X} | D^{(k)} = 1), \quad \{(y_i^{(k)}, \boldsymbol{x}_i^{(k)})\}_{i=1}^{n^{(k)}} \overset{\text{iid}}{\sim} P(Y^{(k)}, \boldsymbol{X}),$$

for $k = 0, 1$. This setting is much easier to apply to real-world situations than the standard setting where the joint samples are required for every individual.

We can interpret this data collection scheme in two ways. First, it can be regarded as a version of *the repeated cross-sectional (RCS) design* (Moretti 2004; Athey and Imbens 2006; Güell and Hu 2006; Ridder and Moffitt 2007; Lebo and Weber 2015) with $k = 0, 1$ representing the time points before and after the assignment regime switches, respectively. The 1E2RD is also possible in this case by setting $P(Z^{(0)} = 1 | \boldsymbol{X}^{(0)}) = 0$. This means collecting data for $k = 0$ at some point before an experiment is conducted. Generally, although panel data are advantageous for statistical analyses because they follow the same individuals over multiple time periods, RCS data have some advantages over panel data. RCS data are easier and cheaper to collect in many cases. Consequently, they are often more representative of the population of interest and larger in sample size than panel data.

However, there is a concern regarding the validity of Assumption 1.1 when we use RCS data since the potential variables may change over time. We need some sort of side information to be confident about the use of RCS data in our setting as we cannot directly test whether the potential variables remain unchanged.

The second interpretation is to collect data with $k = 0, 1$ at the same time by randomly splitting the population of interest. If an experimenter is able to surely perform random splitting, this approach is favorable since we do not have to worry about the validity of Assumption 1.1. The implementation of random splitting is easy in the case of RCTs.

One specific example of our setting is O2O marketing. To measure the effect of an online video ads on sales in physical stores, we usually cannot link the data of those who watched the ads with that of those who shopped at the stores. However, it is relatively easy to separately collect data from those who watch the ads and the purchasers. In this case, data collection can be performed by either RCS design or random splitting.

Another example is when estimating the effect of a treatment that takes a certain period of time to take effect. For instance, it is desirable to use panel data to estimate the effect of job training on future earnings. However, individuals may gradually drop out of the survey, and the probability of the attrition can depend on unobserved variables (Hirano et al. 2001; Nevo 2003). Therefore, it is easier to collect RCS data than to construct balanced panel data. Because we do not need to observe the outcome and treatment status simultaneously, we are more likely to be able to collect even larger and more representative data than with a normal RCS design.

3 Related Works

There have been many proposals for the estimation of $\mu(\boldsymbol{X})$ from the joint samples (Little and Yau 1998; Hirano et al. 2000; Abadie 2003; Tan 2006; Okui et al. 2012; Ogburn, Rotnitzky, and Robins 2015; Wang et al. 2021). These include estimation via the parametric specification of the local average response function $E[Y_d | D_1 > D_0, \boldsymbol{X}]$ (Abadie 2003), doubly robust estimators (Okui et al. 2012; Ogburn, Rotnitzky, and Robins 2015) and estimation with a binary outcome (Wang et al. 2021), to name a few. However, the estimation of $\mu(\boldsymbol{X})$ by data combination has rarely been considered in the literature.

The 2RD has a similar structure to that of *the instrumented difference-in-differences (IDID)* (Ye et al. 2021). The main differences between them are: the conditional independence assumption in IDID is strictly milder than our Assumption 2.1, but IDID requires Z, and the direct estimation is not possible. Which setting is more plausible and practical depends on an actual situation.

Our problem setting is also closely related to *the two-sample IV (TSIV)* estimation (Angrist and Krueger 1992; Inoue and Solon 2010; Pacini and Windmeijer 2016; Choi, Gu, and Shen 2018; Buchinsky, Li, and Liao 2021), where the outcome, IVs, and covariates are not jointly observed in a single dataset. Furthermore, the idea of using moments separately estimated from different datasets can be dated back to at least (Klevmarken 1982). Although estimands in the TSIV estimation are not limited to causal parameters, there have been some studies on the causal inference in the setting related to TSIV. They include ATE estimation from samples of $(Y, Z, \boldsymbol{X})$ and $(D, Z, \boldsymbol{X})$ with the existence of unmeasured confounders (Sun and Miao 2022) and causal inference using samples from heterogeneous populations (Zhao et al. 2019; Shu and Tan 2020). Our setting differs from theirs in that we have to observe D only when $D = 1$, and we do not need Z at all. Particularly, it is of great practical benefit since samples of those who do not receive treatment are often rather difficult to observe. Although our setting requires two regimes, it rather opens up the possibility of LATE estimation in the real world since the 1E2RD is possible.

Another approach for the causal inference from separately observed samples is *the partial identification* (Manski 2003; Tamer 2010; Molinari 2020), that is, deriving bounds for the treatment effects rather than imposing strong assumptions sufficient for the point identification. In (Fan, Sherman, and Shum 2014, 2016), the moment inequality proposed in (Cambanis, Simons, and Stout 1976) was applied to derive the sharp bounds for ATE when only samples of (Y, D) and $(D, \boldsymbol{X})$ are available. The difficulty of applying their approach to the estimation of a treatment effect conditional on covariates is that it requires conditional distribution functions or conditional quantile functions, which are usually difficult to estimate accurately. Moreover, it is sometimes difficult to derive informative bounds for treatment effects without imposing strong assumptions depending on the data generating process (Molinari 2020).

4 Existing Methods

Although the LATE estimation in our specific setting has not been studied before, some existing methods can be applied. We discuss the advantages and disadvantages of these methods especially in terms of accuracy and model selection. Since the true value of treatment effects is not observable by *the fundamental problem of causal inference* (Holland 1986), model selection and hyperparameter tuning are substantial issues in practice (Rolling and Yang 2014; Alaa and van der Schaar 2018; Saito and Yasui 2020).

4.1 Separate Estimation

A naive estimation method for $\mu(\boldsymbol{X})$ is to separately estimate the components and combine them as

$$\widehat{\mu}_{\mathrm{sep}}(\boldsymbol{x}) = \frac{\widehat{E}[Y^{(1)} | \boldsymbol{X} = \boldsymbol{x}] - \widehat{E}[Y^{(0)} | \boldsymbol{X} = \boldsymbol{x}]}{\widehat{E}[D^{(1)} | \boldsymbol{X} = \boldsymbol{x}] - \widehat{E}[D^{(0)} | \boldsymbol{X} = \boldsymbol{x}]}, \quad (2)$$

where a hat denotes an estimator. $E[D^{(k)} | \boldsymbol{X}]$ can be estimated by the PU learning (Elkan and Noto 2008; du Plessis, Niu, and Sugiyama 2014, 2015; Kiryo et al. 2017) with the logistic loss by using $\{\boldsymbol{x}_{di}^{(k)}\}_{i=1}^{n_d^{(k)}}$ as positive data, and the covariates in $\{(y_i^{(1)}, \boldsymbol{x}_i^{(1)})\}_{i=1}^{n^{(1)}}$ and $\{(y_i^{(0)}, \boldsymbol{x}_i^{(0)})\}_{i=1}^{n^{(0)}}$ as unlabeled data. Separate estimation is easy to implement, but usually does not provide a good estimate since it has four possible sources of error.

One advantage of the separate estimation is that the model selection can also be naively performed by choosing the best model for each component. We can easily calculate the model selection criteria, such as the mean squared error (MSE) for each component from the separately observed datasets. However, the resulting $\widehat{\mu}_{\mathrm{sep}}$ may perform poorly because it does not necessarily minimize the model selection criterion in terms of the true μ (Rolling and Yang 2014).

We can alleviate the drawback of the separate estimation by directly estimating the numerator and denominator in (2). Let $T := KD^{(1)} - (1 - K)D^{(0)}$ and $U := KY^{(1)} - (1 - K)Y^{(0)}$ be the auxiliary variables for the notational and computational convenience, and assume $P(K = 1 | \boldsymbol{X}) = 0.5$. Then, we can rewrite the expression in Theorem 1 as $\mu(\boldsymbol{X}) = \nu(\boldsymbol{X})/\pi(\boldsymbol{X})$, where $\nu(\boldsymbol{X}) := E[U | \boldsymbol{X}]$ and $\pi(\boldsymbol{X}) := E[T | \boldsymbol{X}]$. To estimate $\nu(\boldsymbol{X})$ and $\pi(\boldsymbol{X})$, we can construct combined datasets

$$\{(t_i, \boldsymbol{x}_{ti}, r_{ti})\}_{i=1}^{n_t} := \left\{\left((-1)^{1-k}, \boldsymbol{x}_{di}^{(k)}, \frac{\widehat{p}_d^{(k)}(n_d^{(1)} + n_d^{(0)})}{2n_d^{(k)}}\right)\right\}_{i=1,k=0}^{n_d^{(k)},1}$$

$$\{(u_i, \boldsymbol{x}_{ui}, r_{ui})\}_{i=1}^{n_u} := \left\{\left((-1)^{1-k} y_i^{(k)}, \boldsymbol{x}_i^{(k)}, \frac{n^{(1)} + n^{(0)}}{2n^{(k)}}\right)\right\}_{i=1,k=0}^{n^{(k)},1}$$

from $\{\boldsymbol{x}_{di}^{(k)}\}_{i=1}^{n_d^{(k)}}$ and $\{(y_i^{(k)}, \boldsymbol{x}_i^{(k)})\}_{i=1}^{n^{(k)}}$, respectively, where $n_t := n_d^{(1)} + n_d^{(0)}$ and $n_u := n^{(1)} + n^{(0)}$.

We can approximate the expectation of a product of any function of $\boldsymbol{X}$ and T or U by the simple sample average using the above datasets since we have

$$\frac{1}{n_t} \sum_{i=1}^{n_t} r_{ti} t_i f(\boldsymbol{x}_{ti}) = \frac{\widehat{p}_d^{(1)}}{2n_d^{(1)}} \sum_{i=1}^{n_d^{(1)}} f(\boldsymbol{x}_{di}^{(1)}) - \frac{\widehat{p}_d^{(0)}}{2n_d^{(0)}} \sum_{i=1}^{n_d^{(0)}} f(\boldsymbol{x}_{di}^{(0)})$$
$$\approx \frac{1}{2} E[D^{(1)} f(\boldsymbol{X})] - \frac{1}{2} E[D^{(0)} f(\boldsymbol{X})],$$

$$\frac{1}{n_u} \sum_{i=1}^{n_u} r_{ui} u_i f(\boldsymbol{x}_{ui}) = \frac{1}{2n^{(1)}} \sum_{i=1}^{n^{(1)}} y_i f(\boldsymbol{x}_i^{(1)}) - \frac{1}{2n^{(0)}} \sum_{i=1}^{n^{(0)}} y_i f(\boldsymbol{x}_i^{(0)})$$
$$\approx \frac{1}{2} E[Y^{(1)} f(\boldsymbol{X})] - \frac{1}{2} E[Y^{(0)} f(\boldsymbol{X})].$$

An estimator of $\nu(\boldsymbol{X})$ can be obtained by any regression method using the above combined dataset, while we need a little twist to obtain an estimator of the PSD $\pi(\boldsymbol{X})$. See Appendix C in the supplementary material for the direct estimation methods for the PSD.

4.2 Direct Least Squares Estimation

We can apply *the direct least squares (DLS)* method (Yamane et al. 2018) originally proposed for ATE estimation under complete compliance. Motivated by the drawbacks of the separate estimation, the DLS directly estimates $\mu(\boldsymbol{X})$, which is advantageous not only in performance but also in computational efficiency.

The following theorem corresponds to Theorem 1 in (Yamane et al. 2018).

Theorem 2. *Assume $\nu \in L^2$, where $L^2 := \{f : \mathcal{X} \mapsto \mathbb{R} | E[f(\boldsymbol{X})^2] < \infty\}$ in addition to Assumption 1 and 2. Furthermore, define $H_f(\boldsymbol{X}) := \pi(\boldsymbol{X}) f(\boldsymbol{X}) - \nu(\boldsymbol{X})$. Then, μ is equal to the solution of the following least squares problem:*

$$\min_{f \in L^2} E\left[H_f(\boldsymbol{X})^2\right]. \quad (3)$$

Theorem 2 immediately follows from Theorem 1. In what follows, we show that the minimizer of the problem (3) can be estimated without going through the estimation of the conditional mean functions ν and π. Since $(H_f(\boldsymbol{X}) - g(\boldsymbol{X}))^2 \geq 0$ for any square integrable function $g \in L^2$, we have $H_f(\boldsymbol{X})^2 \geq 2H_f(\boldsymbol{X})g(\boldsymbol{X}) - g(\boldsymbol{X})^2$ by expanding the square. The equality holds at $g(\boldsymbol{X}) = H_f(\boldsymbol{X})$ for any f, which maximizes $2H_f(\boldsymbol{X})g(\boldsymbol{X}) - g(\boldsymbol{X})^2$ with respect to g. Hence,

$$\mu(\boldsymbol{X}) = \operatorname*{argmin}_{f \in L^2} \max_{g \in L^2} J(f, g), \quad (4)$$

where $J(f, g) := E\left[2H_f(\boldsymbol{X})g(\boldsymbol{X}) - g(\boldsymbol{X})^2\right]$. Rewriting $J(f, g)$ yields

$$J(f, g) = 2E[Tf(\boldsymbol{X})g(\boldsymbol{X})] - 2E[Ug(\boldsymbol{X})] - E[g(\boldsymbol{X})^2]. \quad (5)$$

While the objective functional in (3) requires the conditional mean estimators, this form can be estimated based on the sample averages as

$$\widehat{J}(f, g) = \frac{2}{n_t} \sum_{i=1}^{n_t} r_{ti} t_i f(\boldsymbol{x}_{ti}) g(\boldsymbol{x}_{ti}) - \frac{2}{n_u} \sum_{i=1}^{n_u} r_{ui} u_i g(\boldsymbol{x}_{ui})$$
$$- \frac{1}{n_u} \sum_{i=1}^{n_u} r_{ui} g(\boldsymbol{x}_{ui})^2.$$

We can implement the DLS estimation of μ with an arbitrary regularization term Ω in practice:

$$\widehat{\mu}_{\mathrm{dls}}(\boldsymbol{x}) = \operatorname*{argmin}_{f \in F} \max_{g \in G} \widehat{J}(f, g) + \Omega(f, g),$$

where F and G are model classes for f and g, respectively.

Although any model can be trained by optimizing the model parameters to minimize the above objective functional, a practically useful choice is a linear-in-parameter model. Set $F = \{f_{\boldsymbol{\alpha}} : \boldsymbol{x} \mapsto \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}) | \boldsymbol{\alpha} \in \mathbb{R}^{q_f}\}$ and $G = \{g_{\boldsymbol{\beta}} : \boldsymbol{x} \mapsto \boldsymbol{\beta}^\top \boldsymbol{\psi}(\boldsymbol{x}) | \boldsymbol{\beta} \in \mathbb{R}^{q_g}\}$, where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are q_f- and q_g-dimensional basis functions, respectively. Using the ℓ_2-regularizer, we have

$$\widehat{J}(f, g) + \Omega(f, g) = 2\boldsymbol{\alpha}^\top \boldsymbol{A} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{b} - \boldsymbol{\beta}^\top \boldsymbol{C} \boldsymbol{\beta}$$
$$+ \lambda_f \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \lambda_g \boldsymbol{\beta}^\top \boldsymbol{\beta},$$

where

$$\boldsymbol{A} := \frac{1}{n_t} \sum_{i=1}^{n_t} r_{ti} t_i \boldsymbol{\phi}(\boldsymbol{x}_{ti}) \boldsymbol{\psi}(\boldsymbol{x}_{ti})^\top, \quad \boldsymbol{b} := \frac{1}{n_u} \sum_{i=1}^{n_u} r_{ui} u_i \boldsymbol{\psi}(\boldsymbol{x}_{ui}),$$

$$\boldsymbol{C} := \frac{1}{n_u} \sum_{i=1}^{n_u} r_{ui} \boldsymbol{\psi}(\boldsymbol{x}_{ui}) \boldsymbol{\psi}(\boldsymbol{x}_{ui})^\top,$$

and λ_f and λ_g are some positive constants. The advantage of this formulation with the linear-in-parameter models and ℓ_2-regularizer is that we have an analytical solution. The solution to the inner maximization is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{C} + \lambda_g \boldsymbol{I}_{q_g})^{-1}(\boldsymbol{A}^\top \boldsymbol{\alpha} - \boldsymbol{b}),$$

where $\boldsymbol{I}_{q_g}$ denotes the $q_g \times q_g$ identity matrix. We can then obtain the DLS estimator of $\boldsymbol{\alpha}$ by substituting $\widehat{\boldsymbol{\beta}}$ and solving the outer minimization:

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{dls}} = \left\{\boldsymbol{A}(\boldsymbol{C} + \lambda_g \boldsymbol{I}_{q_g})^{-1}\boldsymbol{A}^\top + \lambda_f \boldsymbol{I}_{q_f}\right\}^{-1} \boldsymbol{A}(\boldsymbol{C} + \lambda_g \boldsymbol{I}_{q_g})^{-1}\boldsymbol{b}.$$

For the model selection, we can choose a model that minimizes $\widehat{J}(\widehat{f}, \widehat{g})$ evaluated on a validation set. However, as pointed out in (Yamane et al. 2018), one cannot tell if $\widehat{J}(\widehat{f}, \widehat{g})$ is small because $\widehat{f}$ is a good solution to the outer minimization, or $\widehat{g}$ is a poor solution to the inner maximization. For this reason, $\widehat{\mu}_{\mathrm{dls}}(\boldsymbol{x}) := \widehat{\boldsymbol{\alpha}}_{\mathrm{dls}}^\top \boldsymbol{\phi}(\boldsymbol{x})$ can be unstable, and one cannot be confident about which model is the best. The increased dimensionality of the search space because of the need to simultaneously select models for f and g can also make the model selection based on $\widehat{J}(\widehat{f}, \widehat{g})$ challenging.

5 Proposed Method

We propose a novel estimator that enables simpler model selection than the DLS estimation while maintaining the essence of direct estimation as much as possible. We avoid the minimax formulation of the objective functional as in (4) by estimating μ as a solution to the weighted least squares problem derived from the original problem (3). It can be constructed based on the sample averages from the separately observed samples like the DLS estimator, except we need to estimate $\pi(\boldsymbol{X})$ as a weight. We term the proposed estimator as *the directly weighted least squares (DWLS) estimator* since the pre-estimated PSD directly appears without inversion in the objective unlike the IPW estimators (Wooldridge 2002, 2007; Seaman and White 2013).

Shape	n	q_x	DWLS	IWLS	SEP	DLS
h_{con}	10K	1	0.65 ± 1.54	5.04 ± 16.5	1.80 ± 2.38	$\mathbf{0.15 \pm 0.20}$
		5	$\mathbf{0.40 \pm 0.91}$	1.30 ± 3.21	3.46 ± 2.71	0.93 ± 1.17
		10	$\mathbf{0.65 \pm 1.00}$	4.96 ± 14.4	4.18 ± 3.81	1.81 ± 2.82
	50K	1	$\mathbf{0.22 \pm 0.46}$	53.5 ± 372	1.04 ± 0.93	$\mathbf{0.08 \pm 0.13}$
		5	$\mathbf{0.06 \pm 0.07}$	0.35 ± 0.88	2.13 ± 2.00	0.41 ± 0.67
		10	$\mathbf{0.10 \pm 0.11}$	1.22 ± 6.07	3.52 ± 1.92	0.41 ± 0.52
h_{lin}	10K	1	$\mathbf{0.09 \pm 0.15}$	0.51 ± 1.25	0.25 ± 0.30	0.39 ± 0.30
		5	$\mathbf{0.14 \pm 0.08}$	1.32 ± 6.72	0.58 ± 0.41	0.88 ± 0.52
		10	0.61 ± 0.27	4.73 ± 22.4	1.39 ± 0.65	1.53 ± 0.75
	50K	1	$\mathbf{0.02 \pm 0.03}$	6.79 ± 53.1	0.11 ± 0.11	0.36 ± 0.33
		5	$\mathbf{0.04 \pm 0.03}$	0.12 ± 0.16	0.36 ± 0.24	1.04 ± 0.53
		10	$\mathbf{0.32 \pm 0.10}$	11.9 ± 112	0.76 ± 0.26	1.61 ± 0.89
h_{log}	10K	1	$\mathbf{0.19 \pm 0.26}$	0.72 ± 1.82	0.30 ± 0.31	0.32 ± 0.29
		5	$\mathbf{0.29 \pm 0.23}$	12.1 ± 113	0.58 ± 0.40	0.97 ± 0.56
		10	$\mathbf{0.72 \pm 0.31}$	2.76 ± 7.65	1.11 ± 0.49	1.40 ± 0.72
	50K	1	$\mathbf{0.06 \pm 0.08}$	0.30 ± 0.88	0.22 ± 0.16	0.32 ± 0.30
		5	$\mathbf{0.10 \pm 0.04}$	0.18 ± 0.13	0.56 ± 0.31	0.75 ± 0.62
		10	$\mathbf{0.30 \pm 0.10}$	55.3 ± 550	0.86 ± 0.28	1.20 ± 0.99

Table 1: The mean and standard deviation of the MSE over 100 trials. The results are multiplied by 100 (constant), 10 (linear) and 10 (logistic), respectively. The bold face denotes the best and comparative results according to the two-sided Wilcoxon signed-rank test at the significance level of 5%.
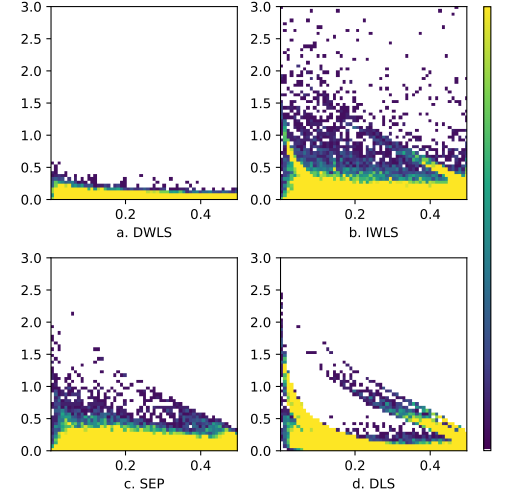

Figure 2: Relation between the squared error (Y-axis) and the PSD (X-axis) when $\mu(\boldsymbol{X})$ is linear, $q_x = 5$ and $n = 10000$. Each point is coloured according to the spatial density.

Consider the following weighted least squares problem:

$$\min_{f \in L^2} E\left[\frac{w(\boldsymbol{X})}{\pi(\boldsymbol{X})} H_f(\boldsymbol{X})^2\right] =: Q_0(f|w), \qquad (6)$$

where $w(\boldsymbol{X})$ is a weight function. Rewriting $Q_0(f|w)$ yields

$$Q_0(f|w) = E[Tw(\boldsymbol{X})f(\boldsymbol{X})^2] - 2E[Uw(\boldsymbol{X})f(\boldsymbol{X})] + E[S], \qquad (7)$$

where $S := \frac{w(\boldsymbol{X})}{\pi(\boldsymbol{X})}\nu(\boldsymbol{X})^2$ is a constant and can therefore be safely ignored in the optimization. Choosing $w = \pi$ clearly reduces the problem (6) to (3). Therefore, we can plug-in the pre-estimated $\pi(\boldsymbol{X})$ as a weight in practice to find a minimizer of the problem (3). However, $\pi(\boldsymbol{X})$ may not be the proper weight since it takes a negative value when $P(Z^{(1)} = 1|\boldsymbol{X}) < P(Z^{(0)} = 1|\boldsymbol{X})$. We can estimate $Q(f|\widehat{\pi}) := Q_0(f|\widehat{\pi}) - E[S]$ based on the sample averages using the separately observed samples as

$$\widehat{Q}(f|\widehat{\pi}) = \frac{1}{n_t}\sum_{i=1}^{n_t} r_{ti} t_i \widehat{\pi}(\boldsymbol{x}_{ti})f(\boldsymbol{x}_{ti})^2 - \frac{2}{n_u}\sum_{i=1}^{n_u} r_{ui} u_i \widehat{\pi}(\boldsymbol{x}_{ui})f(\boldsymbol{x}_{ui}).$$

Using the linear-in-parameter model for f and the ℓ_2-regularizer as in the previous section, we have

$$\widehat{Q}(f|\widehat{\pi}) + \Omega(f) = \boldsymbol{\alpha}^\top \widetilde{\boldsymbol{A}} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \widetilde{\boldsymbol{b}} + \lambda_f \boldsymbol{\alpha}^\top \boldsymbol{\alpha},$$

where

$$\widetilde{\boldsymbol{A}} := \frac{1}{n_t}\sum_{i=1}^{n_t} r_{ti} t_i \widehat{\pi}(\boldsymbol{x}_{ti})\boldsymbol{\phi}(\boldsymbol{x}_{ti})\boldsymbol{\phi}(\boldsymbol{x}_{ti})^\top,$$

$$\widetilde{\boldsymbol{b}} := \frac{1}{n_u}\sum_{i=1}^{n_u} r_{ui} u_i \widehat{\pi}(\boldsymbol{x}_{ui})\boldsymbol{\phi}(\boldsymbol{x}_{ui}).$$

The analytical solution to $\min_{f \in F} \widehat{Q}(f|\widehat{\pi}) + \Omega(f)$ can be obtained as $\widehat{\mu}_{\mathrm{dwls}}(\boldsymbol{x}) := \widehat{\boldsymbol{\alpha}}_{\mathrm{dwls}}^\top \boldsymbol{\phi}(\boldsymbol{x})$, where

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{dwls}} = \left(\widetilde{\boldsymbol{A}} + \lambda_f \boldsymbol{I}_{q_f}\right)^{-1} \widetilde{\boldsymbol{b}}.$$

Although the DWLS objective is no longer the MSE, $\widehat{Q}(\widehat{f}|\widehat{\pi})$ is still a sufficient measure for model selection because $\widehat{\mu}_{\mathrm{dwls}}$ minimizing this objective is also a good estimator in terms of the true MSE as long as $\widehat{\pi}$ is sufficiently accurate. Since the DWLS involves only a single minimization, the model selection is easier than in the DLS estimation.

Remark on the objective formulation We can consider the following least squares problem instead of (6):

$$\min_{f \in L^2} E\left[\frac{\pi(\boldsymbol{X})}{w(\boldsymbol{X})}(f(\boldsymbol{X}) - \mu(\boldsymbol{X}))^2\right] =: Q_0'(f|w). \qquad (8)$$

This objective can also be evaluated without the conditional mean function ν through a transformation similar to (7):

$$Q_0'(f|w) = E\left[\frac{Tf(\boldsymbol{X})^2}{w(\boldsymbol{X})}\right] - 2E\left[\frac{Uf(\boldsymbol{X})}{w(\boldsymbol{X})}\right] + E[S'],$$

where $S' := \frac{\pi(\boldsymbol{X})}{w(\boldsymbol{X})}\mu(\boldsymbol{X})^2$.

We refer to the estimator based on this objective as *the inverse weighted least squares (IWLS) estimator*. This estimator tends to be imprecise when the PSD is extremely close to zero, which is a common issue among the IPW estimators (Wooldridge 2002, 2007; Seaman and White 2013). Although the performance of such estimators can be improved by trimming small probabilities, determining the optimal threshold is nontrivial (Lee, Lessler, and Stuart 2011).

Selecting the threshold is more complicated in the case of the IWLS as the PSD can be negative. At points where the PSD is close to zero, the absolute value of the weights increases while the sign errors are more likely to occur. As a result, a small estimation error in the PSD tends to have a large impact on the IWLS. On the other hand, the impact of the estimation error of the PSD on the DWLS estimator is limited since the absolute value of the weights is small when a sign error occurs in the PSD estimation. The weight

	DR	DWLS	IWLS	SEP	DLS
RMSE	-	252 ± 141	222 ± 145	280 ± 119	946 ± 663
Mean	1778	1528	1559	1538	937
Max	1837	1838	1924	2200	7966
Min	1648	1013	977	919	-1814

Table 2: The results for the performance evaluation using the JTPA dataset. The bold face denotes the best and comparative results according to the two-sided Wilcoxon signed-rank test at the significance level of 5%.

in the proposed method is confined to $[-0.5, 0.5]$, whereas the weight in the IWLS is not bounded at all.

Although the unweighted subsampling method has been recently studied to circumvents weighting samples with the inverse probability (Wang et al. 2020), it cannot be directly applied in this case because it requires the denominator to be a sampling probability.

6 Performance Evaluation

We test the performance of the proposed method with synthetic and real-world datasets. The details of the datasets and other setups can be found in Appendix D in the supplementary material. We denote the separate estimation as the SEP.

6.1 Synthetic Experiment

We generated synthetic data with the different shapes of μ, the covariates' dimension q_x, and the size of training samples. The mean and standard deviation of the MSE over 100 trials are summarized in Table 1. We conducted the Wilcoxon signed-rank test since the IWLS was highly unstable, which made it difficult to use the t-test.

Although we did not trim the PSD for the DWLS estimator, its performance was sufficiently stable to outperform the others in almost all the cases. The DWLS had larger errors than the DLS only when μ was constant and $q_x = 1$. This may be because the benefit of direct estimation outweighed the difficulties of hyperparameter tuning of the DLS since the one-dimensional constant μ is the easiest to learn. This result supports the effectiveness of our approach: avoiding the minimax formulation but without the inverse PSD.

On the other hand, the IWLS estimator often worked terribly poorly even with the weight trimming. The IWLS might be affected by the small PSD most severely since it uses the inverse PSD both in the estimation and hyperparameter tuning. The performance of the DLS was not as poor, but it had a larger MSE than SEP in many cases, indicating that hyperparameter tuning did not work as well as the DWLS.

Figure 2 shows the relation between the squared error of each estimator and the PSD for the linear $\mu(\boldsymbol{X})$ when $q_x = 5$ and $n = 10000$. The area with a dense plot is colored in light. The squared error of the DWLS shown in Figure 2a is kept very small except it is slightly larger when the PSD is close to zero. This result demonstrates the robustness of our method against the near-zero PSD, whereas the performance of the other estimators is affected more severely by the small PSD. We can observe the instability of the IWLS again in Figure 2b, which shows the sporadic high errors over the entire PSD. The squared error of the SEP displays the similar pattern. In Figure 2d, the light-colored area is at

a high position, indicating that the hyperparameter tuning of the DLS does not work well.

6.2 Real Data Analysis

We used the dataset of the National Job Training Partnership Act (JTPA) study[1]. This is one of the largest RCT dataset for job training evaluations in the US with approximately 20000 participants, and it has been used in the several previous studies on causal inference (Bloom et al. 1997; Abadie, Angrist, and Imbens 2002; Donald, Hsu, and Lieli 2014). We used the doubly robust (DR) estimator (Ogburn, Rotnitzky, and Robins 2015) trained with all joint samples as the pseudo true LATE since we do not know the true treatment effect for a real-world dataset.

The results are summarized in Table 2. We report the mean, max, and min of $\widehat{\mu}(\boldsymbol{x})$ in addition to the root mean squared error (RMSE) to see the behavior of each estimator. The IWLS had the smallest RMSE, but the difference from the DWLS was not statistically significant. The mean, max, and min of the DWLS and IWLS also indicate they have similar performance. The stability of the IWLS in contrast to the synthetic experiment is because there were no samples with a small PSD. The relatively good performance of the SEP should be because of the same reason. The performance of the DLS was highly unstable, reflecting the difficulty of the model selection based on the DLS objective.

7 Conclusion

We proposed a novel problem setting for LATE estimation by data combination. Our setting is plausible enough to apply to many real-world problems. The leading examples include O2O marketing and when there is non-random attrition in panel data. We developed a practically useful method for estimating LATE from the separately observed datasets. Our method overcomes the issue of the existing method in model selection by avoiding the minimax formulation of the objective. Furthermore, our method can also avoid the common problem of the IPW methods since the PSD directly appears in our estimator without inversion. Our method displayed the promising performance in the experiments on the synthetic and real-world datasets. However, there are sometimes concerns on the homogeneity of the populations which the separate samples come from. Extending the proposed method to account for heterogeneous samples and time-varying potential variables is an important future direction to further increase its usefulness in practice.

---

[1]https://www.upjohn.org/data-tools/employment-research-data-center/national-jtpa-study

References

Abadie, A. 2003. Semiparametric Instrumental Variable Estimation of Treatment Response Models. *Journal of Econometrics*, 113(2): 231–263.

Abadie, A.; Angrist, J.; and Imbens, G. 2002. Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica*, 70(1): 91–117.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631. Association for Computing Machinery.

Alaa, A.; and van der Schaar, M. 2018. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 129–138. PMLR.

Angrist, J. D.; Imbens, G. W.; and Rubin, D. B. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455.

Angrist, J. D.; and Krueger, A. B. 1992. The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. *Journal of the American Statistical Association*, 87(418): 328–336.

Athey, S.; and Imbens, G. W. 2006. Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica*, 74(2): 431–497.

Bareinboim, E.; and Pearl, J. 2016. Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27): 7345–7352.

Bloom, H. S.; Orr, L. L.; Bell, S. H.; Cave, G.; Doolittle, F.; Lin, W.; and Bos, J. M. 1997. The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study. *The Journal of Human Resources*, 32(3): 549–576.

Buchinsky, M.; Li, F.; and Liao, Z. 2021. Estimation and Inference of Semiparametric Models Using Data from Several Sources. *Journal of Econometrics*.

Cambanis, S.; Simons, G.; and Stout, W. 1976. Inequalities for Ek(X,Y) When the Marginals are Fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 36(4): 285–294.

Choi, J.; Gu, J.; and Shen, S. 2018. Weak-Instrument Robust Inference for Two-Sample Instrumental Variables Regression. *Journal of Applied Econometrics*, 33(1): 109–125.

Donald, S. G.; Hsu, Y.-C.; and Lieli, R. P. 2014. Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT. *Journal of Business & Economic Statistics*, 32(3): 395–415.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2014. Analysis of Learning from Positive and Unlabeled Data. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2015. Convex Formulation for Learning from Positive and Unlabeled Data.

In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, 1386–1394. PMLR.

Elkan, C.; and Noto, K. 2008. Learning Classifiers from Only Positive and Unlabeled Data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220. Association for Computing Machinery. ISBN 9781605581934.

Fan, Y.; Sherman, R.; and Shum, M. 2014. Identifying Treatment Effects Under Data Combination. *Econometrica*, 82(2): 811–822.

Fan, Y.; Sherman, R.; and Shum, M. 2016. Estimation and Inference in an Ecological Inference Model. *Journal of Econometric Methods*, 5(1): 547.

Frölich, M. 2007. Nonparametric IV Estimation of Local Average Treatment Effects with covariates. *Journal of Econometrics*, 139(1): 35–75.

Frölich, M.; and Melly, B. 2013. Identification of Treatment Effects on the Treated with One-Sided Non-Compliance. *Econometric Reviews*, 32(3): 384–414.

Güell, M.; and Hu, L. 2006. Estimating the Probability of Leaving Unemployment Using Uncompleted Spells from Repeated Cross-Section Data. *Journal of Econometrics*, 133(1): 307–341.

Hirano, K.; Imbens, G. W.; Ridder, G.; and Rubin, D. B. 2001. Combining Panel Data Sets with Attrition and Refreshment Samples. *Econometrica*, 69(6): 1645–1659.

Hirano, K.; Imbens, G. W.; Rubin, D. B.; and Zhou, X. H. 2000. Assessing the Effect of an Influenza Vaccine in an Encouragement Design. *Biostatistics*, 1(1): 69–88.

Holland, P. W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396): 945–960.

Imbens, G. W.; and Angrist, J. D. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2): 467–475.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Inoue, A.; and Solon, G. 2010. Two-Sample Instrumental Variables Estimators. *The Review of Economics and Statistics*, 92(3): 557–561.

Kanamori, T.; Hido, S.; and Sugiyama, M. 2009. A Least-Squares Approach to Direct Importance Estimation. *The Journal of Machine Learning Research*, 10: 1391–1445.

Kato, M.; and Teshima, T. 2021. Non-Negative Bregman Divergence Minimization for Deep Direct Density Ratio Estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 5320–5333. PMLR.

Kennedy, E. H. 2020. Efficient Nonparametric Causal Inference with Missing Exposure information. *The International Journal of Biostatistics*, 16(1).

Kiryo, R.; Niu, G.; du Plessis, M. C.; and Sugiyama, M. 2017. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Klevmarken, A. 1982. Missing Variables and Two-Stage Least-Squares Estimation from More than One Data Set. Technical report, IUI Working Paper.

Lebo, M. J.; and Weber, C. 2015. An Effective Approach to the Repeated Cross-Sectional Design. *American Journal of Political Science*, 59(1): 242–258.

Lee, B. K.; Lessler, J.; and Stuart, E. A. 2011. Weight Trimming and Propensity Score Weighting. *PloS one*, 6(3): e18174.

Lee, S.; Correa, J. D.; and Bareinboim, E. 2020. Identifiability from a Combination of Observations and Experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13677–13680. AAAI Press.

Little, R. J.; and Yau, L. H. Y. 1998. Statistical Techniques for Analyzing Data from Prevention Trials: Treatment of No-Shows Using Rubin's Causal Model. *Psychological Methods*, 3(2): 147–159.

Liu, S.; Takeda, A.; Suzuki, T.; and Fukumizu, K. 2017. Trimmed Density Ratio Estimation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Manski, C. F. 2003. *Partial Identification of Probability Distributions*. Springer Science & Business Media.

Molinari, F. 2020. Microeconometrics with Partial Identification. *Handbook of econometrics*, 7: 355–486.

Moretti, E. 2004. Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-Sectional Data. *Journal of Econometrics*, 121(1): 175–212.

Nevo, A. 2003. Using Weights to Adjust for Sample Selection When Auxiliary Information Is Available. *Journal of Business & Economic Statistics*, 21(1): 43–52.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861.

Ogburn, E. L.; Rotnitzky, A.; and Robins, J. M. 2015. Doubly Robust Estimation of the Local Average Treatment Effect Curve. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 77(2): 373.

Okui, R.; Small, D. S.; Tan, Z.; and Robins, J. M. 2012. Doubly Robust Instrumental Variable Regression. *Statistica Sinica*, 22(1).

Oreopoulos, P. 2006. Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter. *American Economic Review*, 96(1): 152–175.

Pacini, D.; and Windmeijer, F. 2016. Robust Inference for the Two-Sample 2SLS Estimator. *Economics letters*, 146: 50–54.

Ridder, G.; and Moffitt, R. 2007. The Econometrics of Data Combination. In Heckman, J. J.; and Leamer, E. E., eds., *Handbook of Econometrics*, volume 6, 5469–5547. Elsevier.

Rolling, C. A.; and Yang, Y. 2014. Model Selection for Estimating Treatment Effects. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 749–769.

Rubin, D. B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5): 688.

Saito, Y.; and Yasui, S. 2020. Counterfactual Cross-Validation: Stable Model Selection Procedure for Causal Inference Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 8398–8407. PMLR.

Seaman, S. R.; and White, I. R. 2013. Review of Inverse Probability Weighting for Dealing with Missing Data. *Statistical Methods in Medical Research*, 22(3): 278–295.

Shu, H.; and Tan, Z. 2020. Improved Methods for Moment Restriction Models with Data Combination and an Application to Two-Sample Instrumental Variable Estimation. *Canadian Journal of Statistics*, 48(2): 259–284.

Skovron, C.; and Titiunik, R. 2015. A Practical Guide to Regression Discontinuity Designs in Political Science. *American Journal of Political Science*, 2015: 1–36.

Sugiyama, M.; Suzuki, T.; and Kanamori, T. 2012. *Density Ratio Estimation in Machine Learning*. Cambridge University Press.

Sun, B.; and Miao, W. 2022. On Semiparametric Instrumental Variable Estimation of Average Treatment Effects through Data Fusion. *Statistica Sinica*. Forthcoming.

Tamer, E. 2010. Partial Identification in Econometrics. *Annual Review of Economics*, 2(1): 167–195.

Tan, Z. 2006. Regression and Weighting Methods for Causal Inference Using Instrumental Variables. *Journal of the American Statistical Association*, 101(476): 1607–1618.

Varian, H. R. 2016. Causal Inference in Economics and Marketing. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27): 7310–7315.

Wang, L.; Zhang, Y.; Richardson, T. S.; and Robins, J. M. 2021. Estimation of Local Treatment Effects under the Binary Instrumental Variable Model. *Biometrika*.

Wang, Z.; Zhu, H.; Dong, Z.; He, X.; and Huang, S.-L. 2020. Less is better: Unweighted data subsampling via influence function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6340–6347. AAAI Press.

Wood, L.; Egger, M.; Gluud, L. L.; Schulz, K. F.; Jüni, P.; Altman, D. G.; Gluud, C.; Martin, R. M.; Wood, A. J.; and Sterne, J. A. 2008. Empirical Evidence of Bias in Treatment Effect Estimates in Controlled Trials with Different Interventions and Outcomes: Meta-Epidemiological Study. *BMJ*, 336(7644): 601–605.

Wooldridge, J. M. 2002. Inverse Probability Weighted M-estimators for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1(2): 117–139.

Wooldridge, J. M. 2007. Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2): 1281–1301.

Yamane, I.; Yger, F.; Atif, J.; and Sugiyama, M. 2018. Uplift Modeling from Separate Labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ye, T.; Ertefaie, A.; Flory, J.; Hennessy, S.; and Small, D. S. 2021. Instrumented Difference-in-Differences. arXiv:2011.03593.

Yeo, I. K.; and Johnson, R. A. 2000. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, 87(4): 954–959.

Zhao, Q.; Wang, J.; Spiller, W.; Bowden, J.; and Small, D. S. 2019. Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples. *Statistical Science*, 34(2): 317–333.