Summary: Identification and estimation of the joint distribution of potential outcomes in causal inference

Keisuke Takahata

Graduate School of Economics, Keio University

1 Chapter 1: Introduction

1.1 Identification problem in econometrics: Motivation and purpose of the thesis

In research of econometric or statistical methodologies, the primal focus is often made on estimation. A typical statistical research mainly considers constructing an estimator and analyzing its asymptotic properties. It enables us to carry out a statistical inference, that is, hypothesis tests and calculating confidence intervals, and also provides a reasonable explanation why a proposed estimator is reasonable in some respects. Besides, the development of a specific algorithm that approximates the estimator of interest is also a core of modern econometric and statistical research. As a sample size is getting larger along with the rapid development of the internet and storage services, and powerful computational resources have become available at an individual level, efficient treatments of estimation with large data are becoming a more and more important than ever.

In those efforts, the problem is how to deal with a finite number of observations to draw inference about a population of interest, and one often implicitly assumes that if one were to obtain an infinite number of observations the problem is solved. In basic statistical problems, this view may not cause a fundamental problem. Classical statistical frameworks were originally developed with agricultural applications in mind. Fisher and Neyman-Pearson frameworks mainly focus on which fertilizer contributes to the growth of crops more effectively. In such studies, a researcher could completely randomize the assignment of a treatment,

and almost fully obtain covariates which are needed in an analysis. Therefore, a researcher has an "ideal" data (compared to econometric applications), and then the problem is how to draw a statistically reasonable conclusion given a limited number of observations. In the last decade, machine learning and deep learning research have made large successes mainly in image recognition and related topics, and been shown to outperform human performances in several specific areas. Although there still remain many technical difficulties to overcome, in principle, they are same as classical statistics in a sense that researchers do not pay much attention to sampling processes nor model identification; predictive performance is of primal interest.

However, in econometrics, those views are often not true. First, when dealing with economic agents, researchers often cannot carry out a randomized experiment in which they assign particular individuals to take or do something and others not to. Besides, even if they can randomize an assignment, it is often the case that not all the individuals in an experiment comply the assignment they receive. In such a situation, the observed data does not represent a population of interest. Second, in observational studies, endogeneity has to be considered. Other than choice of individuals, simultaneity, that is, the result of economic equilibrium, is also an important element that causes endogeneity. Under the endogeneity of a treatment, the observed data is biased, so even if an infinite number of observations from such sampling process is available we cannot be able to draw a correct conclusion about a population of interest unless an additional assumption is introduced. Therefore, in econometric research, one always has to ask oneself whether data at one's hand can potentially reveal a parameter of interest, and if not, what kind of assumptions need to be combined.

To make the issue distinct we should separate an inferential problem into two components: statistical inference and identification components. Studies of identification concern the conclusions that could be obtained if we could obtain an unlimited number of observations. On the other hand, studies of statistical inference consider evaluating generally weaker conclusion about a population arising from a finite number of observations, such as sampling error or rate of convergence. Obviously, if we cannot obtain a conclusion about a parameter given an infinite number of observations, nor can we given a finite number of observations. Therefore, the study of identification has to come first; any statistical study makes no sense until we guarantee the identification of a parameter of interest.

In this thesis, I study identification and estimation of the joint distribution of potential outcomes in causal inference. In a standard causal analysis, a parameter of interest is often an average treatment effect (ATE), or in recent years, heterogeneous treatment effect (HTE), which is a conditional treatment effect for a subset of observed covariates. For identifying these parameters, unconfoundedness plays a primal role. However, if one wants to identify, for example, the correlation between potential outcomes or quantile of individual treatment effects,

unconfoundedness is not sufficient. For they are a functional of the joint distribution of potential outcomes, not of the (conditional) expectations of each potential outcome. Therefore, for valid inference of these parameters, we need to study the identification of the joint distribution of potential outcomes. In subsequent chapters, I study this issue and propose an estimation method for several different setups.

1.2 Basic concepts of identification problems

For a general discussion, here we provide several important concepts and definitions regarding to identification problem based on probability density distribution. Followings are mainly based on Rothenberg (1971).

Let $Y \in \mathbb{R}^d$ be a vector-valued i.i.d. random variable. Suppose that the probability density function of Y belongs to a family $\mathscr{F}$ of density functions on $\mathbb{R}^d$. Suppose that each member of $\mathscr{F}$ is characterized by a parameter α in a parameter space A. Combining the space of $\mathscr{F}$ and A, we denote $\mathscr{F}(A)$ and call it as a *model*. A parameter α uniquely determines a density function $f(\cdot;\alpha)$ in $\mathscr{F}$. In general, α can be infinite-dimensional, but here we focus our attention to finite-dimensional cases for brevity of explanation.

As described before, identification implies that sufficiently large number of observations can distinguish one point from the others in a parameter space. To express this notion, we introduce the following concept:

Definition 1. *Two parameter points α_1 and α_2 are said to be observationally equivalent if $f(y,\alpha_1) = f(y,\alpha_2)$ for all $y \in \mathbb{R}^d$.*

Identification of parametric models is defined through the concept of observationally equivalent as follows:

Definition 2. *A model $\mathscr{F}(A)$ is said to be identifiable or identified if for any $\alpha \in A$ there is no other point in A which is observationally equivalent.*

In mathematical point of view, identifiability is also characterized as the injectivity of density function f:

Definition 3. *A model $\mathscr{F}(A)$ is identifiable or identified if*

$$f(y;\alpha_1) = f(y;\alpha_2) \Rightarrow \alpha_1 = \alpha_2$$

for any $\alpha_1, \alpha_2 \in A$ and any $y \in \mathbb{R}^d$, that is, $f(\cdot;\alpha)$ is injective with respect to α.

The injectivity of a density function will play a significant role in our identification analysis in subsequent chapters.

Note that there is another route in which one allows an identified parameter to be a set, not a point in a parameter space, for avoiding additional restrictions. This approach is called *set identification* or *partial identification* (Manski, 2007; Tamer, 2010), while identification under the definitions provided above are called point identification compared with this, and has been extensively studied in recent years. But we restrict our attention to point identification in the thesis.

Note also that the terms "identifiable" and "identified", or "identifiability" and "identification", are used as synonyms in identification literature (Aldrich, 2002). We use them exchangeably in the thesis.

2 Chapter 2: Parametric Identification of the Joint Distribution of the Potential Outcomes

2.1 Introduction

A typical quantity of interest in causal inference is an average treatment effect (ATE), which is defined as the expectation of the difference of the potential outcomes. Because the potential outcomes are never observed simultaneously for each unit (which is often mentioned as "the fundamental problem of causal inference," (Holland, 1986) one cannot simply take the mean of the difference of the observed outcomes to estimate ATE. Then, certain conditions for identifiability, such as strong ignorability (Rosenbaum and Rubin, 1983) or the instrumental variable approach, have been employed to estimate ATE. ATE on the treated (ATT), which is also a parameter of interest, can be estimated in a similar approach. Identification and estimation of ATE and ATT are based on the fact that they are expressed as the function of the marginal expectations of each potential outcome; then we do not have to identify the joint distribution of the potential outcomes and can estimate them by using observed data from randomized experiments or observational studies under the ignorability condition.

However, there are several parameters of interest in which we cannot rely on this approach for inference. For example, a quantile treatment effect (QTE), a quantile of the difference of the potential outcomes, cannot be estimated correctly even in a randomized experiment due to the fundamental problem of causal inference. Firpo (2007) proposed estimating QTE by the difference of the quantiles of the potential outcomes. However, because taking a quantile is not a linear operator, it is not a consistent estimator for the quantile of the treatment effect in general; we actually need to know the joint distribution of the potential outcomes.

In this chapter, we propose a parametric identification approach for the joint distribution of the potential outcomes. We show that the non-normality of the

distribution of the untreated outcome or the conditional distribution of the treated outcome given the untreated outcome can play a significant role for the identification. Because non-normal distributions are accurately approximated by a mixture of normal distributions with a finite number of components, our result implies the identification of sufficiently large classes of distributions (Ferguson, 1973; Ishwaran and Zarepour, 2000; Ishwaran and James, 2001). We mainly focus on the case of a randomized control trial, but by incorporating covariates into the model, our approach can be extended to an observational study case in which the ignorability assumption holds.

2.2 Main Results

We take the setup of Rubin's potential outcome model (Rubin, 1974). Let $y_0, y_1 \in \mathbb{R}$ be scalar potential outcomes when receiving the control and treatment condition, respectively. Denote by $\mathscr{P}_v(\omega) = \{p(v;\omega) : \omega \in \Omega\}$ the class of the density function of random variable v characterized by the parameter ω. Throughout this chapter, we restrict our attention to continuous potential outcomes.

Let $p(y_0) \in \mathscr{P}_{y_0}(\lambda)$, $p(y_1 \mid y_0) \in \mathscr{P}_{y_1|y_0}(\theta)$, and $p(y_1) \in \mathscr{P}_{y_1}(\xi)$, where $\lambda \in \Lambda$, $\theta \in \Theta$, and $\xi \in \Xi$. Considering the identity

$$p(y_1) = \int p(y_1 \mid y_0) p(y_0) dy_0, \tag{1}$$

we can write

$$\xi = (\xi_1, \xi_2) = (\psi(\theta, \lambda), \overline{\psi}(\lambda)), \tag{2}$$

where $\psi : \Theta \times \Lambda \to \Xi_1 \subset \Xi$, $\overline{\psi} : \Lambda \to \Xi_2 \subset \Xi$, $\Xi_1 \cap \Xi_2 = \emptyset$, $\Xi_1 \cup \Xi_2 = \Xi$. The notation (2) implies that the parameter for $p(y_1)$ can be expressed as the function of that of $p(y_0)$ and $p(y_1 \mid y_0)$, and is partitioned into two parts, one of which is dependent on θ, while the other is independent of θ. Because we can identify the marginal distributions of each potential outcome, that is, λ and ξ, using the observed data, our problem reduces to: under what conditions θ is identified given the information λ and ξ_1. In what follows, we treat λ as a known constant, not a parameter, since our interest lies in the identification of θ.

With this setup, if

$$\psi(\theta, \lambda) = \psi(\theta', \lambda) \Rightarrow \theta = \theta' \tag{3}$$

for any fixed λ, then θ is identified. Therefore, a general identification condition for θ can be stated as follows.

Proposition 1. *The parameter for $p(y_1 \mid y_0)$, θ, is identified if ψ is injective with respect to θ.*

A necessary condition for the injectivity of ψ is

$$\dim(\psi(\theta,\lambda)) \geq \dim(\theta). \tag{4}$$

Then, our identification analysis consists of two steps: (a) first, we examine condition (4), and if it is true, then (b) we try to construct the inverse mapping ν such that, for any fixed λ,

$$\nu \circ \psi(\theta,\lambda) = \theta. \tag{5}$$

Investigating those conditions, we obtain the two main theorems. First one is about a simple normal model.

Theorem 1. *Suppose $p(y_1 \mid y_0) = N(y_1; \beta_0 + \beta_1 y_0, \sigma_{10}^2)$ and that $p(y_0)$ is not a degenerate distribution. Then, the dimension condition (4) is satisfied if and only if $p(y_0)$ is not a normal distribution.*

This theorem suggests that, when the relation between the potential outcomes is supposed to be linear, we may have the identification of the joint distribution of potential outcomes if the distributions of observed variables are skewed.

Second, we extend this result to a more flexible model where the joint distribution $p(y_1, y_0)$ is directly specified by normal mixtures. It is well known that any continuous distribution is represented as the mixture of an infinite number of distributions, except for several unusual cases (Ferguson, 1973). Besides, even with a finite number of components, a mixture distribution provides a good approximation of a target distribution if it has enough components (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001). Then, the specification of $p(y_1, y_0)$ with a finite normal mixture broadens the scope of identifiable relationships between the potential outcomes.

We focus on a model of the form

$$\begin{aligned} p(y_1, y_0) = \sum_{k=1}^{K} w_{(k)} \{ & N(y_1; \beta_{0(k)} + \rho_{(k)} r_{(k)} y_0, (1-\rho_{(k)}^2) r_{(k)}^2 \sigma_{0(k)}^2) \\ & \times N(y_0; \mu_{0(k)}, \sigma_{0(k)}^2) \} \end{aligned} \tag{6}$$

where $\beta_{0(k)} = \mu_{1(k)} - \rho_{(k)} r_{(k)} \mu_{0(k)}$ and $r_{(k)} = \sigma_{1(k)} / \sigma_{0(k)}$. Similar to Theorem 1, we have the following result.

Theorem 2. *Suppose the model (6) ($K \geq 2$) and if:*

(A1) $\rho_{(k)} = \rho$ for $k = 1, \ldots, K$;

(A2) there exists k_1, k_2 such that $\beta_{0(k_1)} = \beta_{0(k_2)}$;

(A3) $w_{(k)} \neq w_{(k')}$ for all $k \neq k'$,

then $p(y_1, y_0)$ is identified.

Note that under (6), the conditional expectation of y_1 given y_0 can be expressed as a simplified form of a local linear regression (Müller et al., 1996):

$$\mathbb{E}[y_1 \mid y_0] = \frac{1}{p(y_0)} \int y_1 p(y_1, y_0) dy_1 = \sum_{k=1}^{K} \overline{w}_{(k)} (\beta_{0(k)} + \rho r_{(k)} y_0),$$

where $\overline{w}_{(k)} = w_{(k)} N(y_0; \mu_{0(k)}, \sigma^2_{0(k)}) / p(y_0)$. Then, even under the identification conditions (A1)–(A3), the model (6) can express complex relationships between the potential outcomes to some extent. Furthermore, we can introduce covariates to (6), which enables our analysis to be extended to observational studies under unconfoundedness.

3 Chapter 3: Estimation of Heterogeneous Treatment Effects under Non-ignorable Assignment

3.1 Introduction

Extending the methods developed in Chapter 2, we consider identifying and estimating heterogeneous treatment effects (HTEs) in non-randomized settings. Existing studies often define HTEs as a function of the observable variables; however in this chapter we discuss the identification and inference of HTEs which we define as

$$\mathrm{HTE}(y_0) = \mathbb{E}[y_1 - y_0 \mid y_0],$$

where $y_1 \in \mathbb{R}$ and $y_0 \in \mathbb{R}$ are the potential outcome variable under the (special) treatment condition (with higher cost) and the (default) control condition respectively. This HTE is a function of y_0, which can indicate how much effect the unit whose outcome is y_0 under the untreated condition, would get if the unit is assigned to the treatment condition. This information may be more informative to policy makers or medical practitioners than HTEs in a standard definition.

To this end, we introduce the following two assumptions. First, we consider relaxing strong ignorability condition, which is often assumed in observational studies, as

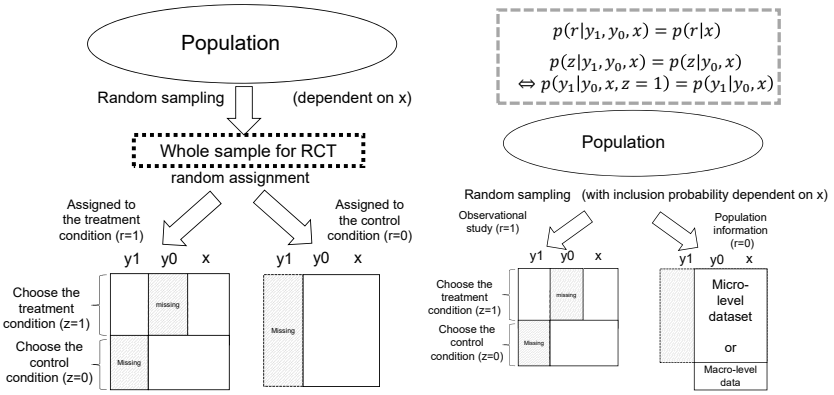
$$p(z \mid y_1, y_0, x) = p(z \mid y_0, x), \tag{7}$$

where $x \in \mathbb{R}^d$ is a d-dimensional covariate vector and $z \in \{0, 1\}$ is the binary indicator which is $z = 1$ when y_1 is observed (i.e. when assigned and complying with the treatment condition). We refer to this assumption as *weak ignorability*.

Besides, although we assumed completely randomized experiments in the last chapter, here we extend them to a case in which there exists a nonignorable

noncompliance under weak ignorability. Under weak ignorability, the probability of being treated is not identified in general. To deal with this issue, we assume that the information on the distribution of the untreated outcome $p(y_0)$ or its moments is available.

As typical examples where $p(y_0)$ can be obtained, we consider the following two setups prevalent in applied studies (see also Figure 1): (a) randomized controlled trials which "one-sided noncompliance" (Imbens and Rubin, 2015), in that for the control group all the participants comply with the control condition while for the treatment group not all the participants comply with their treatment, or individuals are allowed to choose their treatment, and (b) observational studies in which external information on the population or a random sample of the population is available.



(a) RCT with one-sided noncompliance

(b) Observational study with population information

Figure 1: The two setups we considered in this chapter

3.2 Main results

Applying the result of Hirano et al. (2001), who consider the identification of a model for a panel data in which there is a non-ignorabile attrition but refreshment samples avaiable, and Takahata and Hoshino (2019), we have the following main theorem.

Theorem 2. *Under weak ignorability and if:*

(c.1) $p(y_0)$ *is known;*

(c.2) $p(z = 1 \mid y_0, x)$ *has no interaction term between* y_0 *and* x*;*

(c.3) $y_1 = \gamma_0 y_0 + \phi_{10}(x) + \varepsilon_{10}$, $\varepsilon_{10} \sim N(0, \sigma_{10}^2)$, *where* $\phi(x)$ *is an arbitrary function of* x*;*

(c.4) $p(y_0 \mid x, z = 1)$ *is not a degenerate distribution nor the normal distribution nor a mixture of them;*

then, HTE and ATE are identifiable.

Condition (c.3) requires that the conditional expectation of y_1 given y_0 and x is linear with respect to y_0. This seems to be somewhat restrictive, but the conditional expectation of y_1 given y_0 is generally nonlinear with respect to y_0 by integrating out x. Therefore, various kinds of heterogeneity in treatment effects can be captured to some extent even under (c.3).

Based on Theorem 2, we propose a quasi-Bayesian estimation method. To reflect an auxiliary information in the likelihood, we rely on a GMM-based method by Nevo (2003). Combing this method with the data augmentation algorithm (or numerical integration), we can derive a MCMC procedure for posterior approximation.

We apply the proposed method to real job-training data set from the National Job Training Partnership Act (JTPA) Study. This data set has been analyzed by various researchers. Based on the estimates of HTE, we show a novel findings about women participants, along with consistent result with existing studies.

4 Chapter 4: Nonparametric point identification and a Bayesian estimation of the joint distribution of potential outcomes

4.1 Introduction

This chapter further extend the results in previous chapters from parametric specifications to nonparametric ones.

In economic program evaluation, average treatment effects (ATEs) has long been a primal parameter of interest. Identification and estimation of ATEs are extensively studied (see Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018) for a review). While ATEs summarize the effect of a treatment of interest by a scaler value, distributional treatment effects are known to be more informative than ATEs in some applications.

Quantile treatment effects (QTEs) are a special case of distributional treatment effect and many methods are suggested for the identification and estimation of QTEs (e.g. Chernozhukov and Hansen (2005) and Firpo (2007)). However, QTEs considered in previous studies are not a quantile of a treatment effect, as discussed in Chapter 2. To deal with inferences on exactly defined QTEs, we

need to know the joint distribution of the potential outcomes. In addition to QTEs, there are several parameters in which we need to identify the joint distribution of the potential outcomes. Takahata and Hoshino (2019) discuss several examples, including causal mediation analysis and heterogeneous treatment effects.

To overcome this issues, we consider a sufficient condition for nonparametric point identification of the joint distribution of the potential outcomes. As described later, the identification problem reduces to the uniqueness of the solution to certain integral equation. Hu et al. (2017) illustrate that several economic problems which concern the uniqueness of an integral equation can be solved by transforming it to a Volterra integral equation of the second-kind. They propose the moving support condition, which requires the support of one of the potential outcomes to vary according to the control variable, to make the transormation possible. Volterra integral equations of the second-kind are known to have a unique solution, then we readily obtain the identification once the integral equation of interest is transformed to it.

For this purpose, we assume the existence of a control variable, which is employed in various literature in econometric research. We show that, given the existence of the control variable which satisfy the moving support condition, the joint distribution of the potential outcomes is nonparametrically identified. While the existence of the control variable is untestable and its validity needs to be argued based on the background of a specific example, we demonstrate that the moving support condition is approximately satisfied in the wide range of distributions. For estimation, we employ a nonparametric Bayesian density estimation with the data augmentation technique to deal with the missing outcomes. Finally, we show the results of numerical experiments that encourage our identification result.

4.2 Main results

We follow Rubin's potential outcome framework (Rubin, 1974, 1990; Imbens and Rubin, 2015), as same as the previous chapters. We focus on randomized experiments without non-compliance, but the analysis may be extended to observational studies or other designs by introducing additional ignorability conditions.

Suppose that a vector of pretreatment covariates $w \in \mathbb{R}^d$ is available for all the individuals. The key identity is

$$p(y_1 \mid w) = \int p(y_1 \mid y_0, w) p(y_0 \mid w) dy_0. \tag{8}$$

If all the variables are discrete, we consider the integral in (8) as summation. In a randomized experiment, $p(y_1 \mid w)$ and $p(y_0 \mid w)$ are identified using the observations. Therefore, we can consider the identity (8) as an integral equation with

respect to $p(y_1 \mid y_0, w)$ and the identification problem reduces to the uniqueness of the solution to (8). Note that the whole discussion of this chapter holds if we switch roles between y_1 and y_0. That is, we can also develop our analysis on the identity

$$p(y_0 \mid w) = \int p(y_0 \mid y_1, w) p(y_1 \mid w) dy_1,$$

instead of (8). However, for simplicity of exposition, we develop our analysis based on (8).

In the field of integral equations a known function inside the integral is called a *kernel.* In our case, the kernel is $p(y_0 \mid w)$. The difficulty of our problem is that the dimension of arguments of the kernel is smaller than that of the function of interest. Integral equations of such class have not been studied well, and do not allow us to investigate theoretical properties. Therefore, we introduce several assumptions that enable us to refer to the existing framework of integral equations.

The most significant assumption made in our analysis is the following.

Assumption 1. *(existence of a control variable) There exists a subset of* $w = (v, u)$ *such that*

$$\begin{aligned} &y_0 \not\perp\!\!\!\perp v \mid u, \\ &y_1 \perp\!\!\!\perp v \mid y_0, u. \end{aligned} \tag{9}$$

Intuitively, this requires that there exists a variable v such that it does not affect a heterogeneity of treatment effects. Assumption 1 is related to the *control function assumption*, which is employed in various kinds of literature, such as (nonparametric) simultaneous equation models, nonseparable models, or measurement error models (Heckman and Robb Jr, 1985; Newey et al., 1999; Blundell and Powell, 2003; Florens et al., 2008; Su and Ullah, 2008; Imbens and Newey, 2009; Blundell et al., 2013; Wiesenfarth et al., 2014).

Under Assumption 1, we have

$$p(y_1 \mid v, u) = \int p(y_1 \mid y_0, u) p(y_0 \mid v, u) dy_0. \tag{10}$$

By fixing y_1 at an arbitrary (finite) value, we observe that the kernel $p(y_0 \mid v, u)$ has the arguments of larger dimensions than the function of interest. Therefore, we can study the uniqueness of the solution to (9) based on the existing literature on linear integral equations.

Without loss of generality, we can assume that v is a scalar. Besides, the discussions below hold for any fixed u. Then, we will suppress the conditioning on u in what follows and rewrite the integral equation of interest as

$$p(y_1 \mid v) = \int p(y_1 \mid y_0) p(y_0 \mid v) dy_0. \tag{11}$$

It is still difficult for us to analyze the identification condition with this expression. However, if we consider (11) as a Volterra integral equation of the first-kind, we can develop further our analysis.

Hu et al. (2017) suggests a set of regularity conditions that a Volterra equation of the first-kind can be transformed into the second-kind, which is known to admit a unique solution. Here we adapt their approach to our setting. Suppose that the range of integration in (11) is bounded, and is a function of v, $[\underline{b}(v), \overline{b}(v)]$, which is called the *moving support condition*. Fixing y_1 and viewing it as a "parameter", we rewrite (11) as

$$h(v) = \int_{\underline{b}(v)}^{\overline{b}(v)} K(v, y_0) g(y_0) dy_0, \tag{12}$$

where, $h(v) = p(y_1 \mid v)$, $K(v, y_0) = p(y_0 \mid v)$, and $g(y_0) = p(y_1 \mid y_0)$.

Similar to Theorem 2.1 in Hu et al. (2017), we obtain the identification of $g(y_0)$, that is, $p(y_1 \mid y_0)$ for fixed y_1. Since this argument holds for arbitrarily fixed y_1, we obtain the identification of $p(y_1 \mid y_0)$.

Theorem 3. *Suppose Assumptions 1–8 (see the main text) hold. Then, given the identification of $p(y_1 \mid v)$ and $p(y_0 \mid v)$, $p(y_1 \mid y_0)$ is identified.*

The moving support condition and associated regularity conditions look difficult to be satisfied. Particularly, that the upper bound of y_0 must be monotone increasing and the support diminish on the lower bound seems to be demanding. However, we can demonstrate that it can be approximately satisfied in many cases when we transform variables each of which is supported on a real line into the unit interval, $[0, 1]$. Importantly, several standard specifications on $\mathbb{R}^3$ approximately satisfy the moving support condition by this transformation, while there are some exceptions.

To examine the derived identification results, we propose a Bayesian nonparametric estimation method and carry out numerical experiments based on it. In frequentist's approach, we need to integrate out the unobserved variables (i.e., y_0 in the treatment group and y_1 in the control group) in the conditional likelihood function for inference. With a large sample, it requires a numerical integration as high-dimensional as the sample size, which is computationally expensive. On the other hand, in Bayesian inference, we can use data augmentation algorithms (Tanner and Wong, 1987; Damien et al., 1999), where the integration for unobserved variables is replaced by simply sampling unobserved outcomes given the other parameters in each iteration of Markov Chain Monte Carlo algorithm, and marginalizing these augmented variables. Although its computational cost also increases with the sample size, sampling is by far simpler than numerical integration. In this respect, the use of a Bayesian approach is encouraging.

We employ Dirichlet Process Mixture models (Ferguson, 1973; MacEachern and Müller, 1998; Ishwaran and James, 2001; Gelman et al., 2013) for avoiding

misspecification and taking advantage of the nonparametric identification result. Given the augmented variables, the Gibbs sampling algorithm based on the stick-breaking prior can be used, which allows for simple posterior computations.

The results of experiments for several different data-generating processes encourage our identification results.

References

ABADIE, A. AND M. D. CATTANEO (2018): "Econometric methods for program evaluation," *Annual Review of Economics*, 10, 465–503.

ALDRICH, J. (2002): "How likelihood and identification went Bayesian," *International Statistical Review*, 70, 79–98.

BLUNDELL, R., D. KRISTENSEN, AND R. L. MATZKIN (2013): "Control functions and simultaneous equations methods," *American Economic Review*, 103, 563–69.

BLUNDELL, R. AND J. L. POWELL (2003): "Endogeneity in nonparametric and semiparametric regression models," *Econometric society monographs*, 36, 312–357.

CHERNOZHUKOV, V. AND C. HANSEN (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261.

DAMIEN, P., J. WAKEFIELD, AND S. WALKER (1999): "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables," *Journal of the Royal Statistical Society, Series B*, 61, 331–344.

FERGUSON, T. (1973): "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, 1, 209–230.

FIRPO, S. (2007): "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75, 259–276.

FLORENS, J.-P., J. J. HECKMAN, C. MEGHIR, AND E. VYTLACIL (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191–1206.

GELMAN, A., J. B. CARLIN, H. S. STERN, D. B. DUNSON, A. VEHTARI, AND D. B. RUBIN (2013): *Bayesian Data Analysis*, Chapman and Hall/CRC, 3rd ed.

HECKMAN, J. J. AND R. ROBB JR (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of econometrics*, 30, 239–267.

HIRANO, K., G. W. IMBENS, G. RIDDER, AND D. B. RUBIN (2001): "Combining panel data sets with attrition and refreshment samples," *Econometrica*, 69, 1645–1659.

HOLLAND, P. W. (1986): "Statistics and causal inference," *Journal of the American Statistical Association*, 81, 945–960.

HU, Y., S. M. SCHENNACH, AND J.-L. SHIU (2017): "Injectivity of a class of integral operators with compactly supported kernels," *Journal of Econometrics*, 200, 48–58.

IMBENS, G. W. AND W. K. NEWEY (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

IMBENS, G. W. AND D. B. RUBIN (2015): *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

IMBENS, G. W. AND J. M. WOOLDRIDGE (2009): "Recent developments in the econometrics of program evaluation," *Journal of economic literature*, 47, 5–86.

ISHWARAN, H. AND L. F. JAMES (2001): "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

ISHWARAN, H. AND M. ZAREPOUR (2000): "Markov chain monte carlo in approximate Dirichlet and beta two-parameter process hierarchical models," *Biometrika*, 87, 371–390.

MACEACHERN, S. N. AND P. MÜLLER (1998): "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

MANSKI, C. F. (2007): *Identification for Prediction and Decision*, Harvard University Press.

MÜLLER, P., A. ERKANLI, AND M. WEST (1996): "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 1221–1246.

NEVO, A. (2003): "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business & Economic Statistics*, 21, 43–52.

NEWEY, W. K., J. L. POWELL, AND F. VELLA (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67, 565–603.

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

ROTHENBERG, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577–591.

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

——— (1990): "Formal mode of statistical inference for causal effects," *Journal of statistical planning and inference*, 25, 279–292.

SU, L. AND A. ULLAH (2008): "Local polynomial estimation of nonparametric simultaneous equations models," *Journal of Econometrics*, 144, 193–218.

TAKAHATA, K. AND T. HOSHINO (2019): "Parametric identification of the joint distribution of the potential outcomes," *Stat*, in press.

TAMER, E. (2010): "Partial identification in econometrics," *Annual Review of Economics*, 2, 167–195.

TANNER, M. A. AND W. H. WONG (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

WIESENFARTH, M., C. M. HISGEN, T. KNEIB, AND C. CADARSO-SUAREZ (2014): "Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures," *Journal of Business & Economic Statistics*, 32, 468–482.