

# ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

*An International Society for the Advancement of Economic  
Theory in its Relation to Statistics and Mathematics*

<http://www.econometricsociety.org/>

*Econometrica*, Vol. 82, No. 2 (March, 2014), 811–822

## IDENTIFYING TREATMENT EFFECTS UNDER DATA COMBINATION

YANQIN FAN

*University of Washington, Seattle, WA 98195, U.S.A.*

ROBERT SHERMAN

*HSS, Caltech, Pasadena, CA 91125, U.S.A.*

MATTHEW SHUM

*HSS, Caltech, Pasadena, CA 91125, U.S.A.*

---

The copyright to this Article is held by the Econometric Society. It may be downloaded, printed and reproduced only for educational or research purposes, including use in course packs. No downloading or copying may be done for any commercial purpose without the explicit permission of the Econometric Society. For such commercial purposes contact the Office of the Econometric Society (contact information may be found at the website <http://www.econometricsociety.org> or in the back cover of *Econometrica*). This statement must be included on all copies of this Article that are made available electronically or in any other format.

---

## IDENTIFYING TREATMENT EFFECTS UNDER DATA COMBINATION

BY YANQIN FAN, ROBERT SHERMAN, AND MATTHEW SHUM<sup>1</sup>

We consider the identification of counterfactual distributions and treatment effects when the outcome variables and conditioning covariates are observed in separate data sets. Under the standard selection on observables assumption, the counterfactual distributions and treatment effect parameters are no longer point identified. However, applying the classical monotone rearrangement inequality, we derive sharp bounds on the counterfactual distributions and policy parameters of interest.

KEYWORDS: Counterfactual distributions, treatment effects, partial identification.

### 1. INTRODUCTION

IN THIS NOTE, we consider how to identify counterfactual distributions and treatment effects when the outcome variables and the conditioning covariates are observed in separate data sets. The need to combine variables from separate data sets arises naturally in many policy applications; these include poverty analysis in which one data set consists of program participation and the other consists of demographic attributes, or epidemiological studies in which incidence of the disease and demographic variables are observed separately.

We consider the identification of counterfactual distributions and treatment effects under the standard unconfoundedness or selection on observables assumption. It is composed of (i) the conditional independence assumption—that is, the potential outcomes are jointly independent of the treatment conditional on a set of observed covariates, and (ii) the common support assumption—that is, the propensity score is strictly between 0 and 1 for all values of the conditioning covariates. When the treatment outcomes and covariates are observed in a single data set, it is well known that the marginal and counterfactual distributions of potential outcomes (and hence the average treatment effects and treatment effects for the treated) are point identified. A voluminous literature has explored many aspects of identification, inference, and computation.<sup>2</sup>

When outcomes and conditioning covariates are observed in separate data sets, the aforementioned point identification results break down. Using explicit

<sup>1</sup>We are grateful to Cheng Hsiao, Sergio Firpo, Marc Henry, Chuck Manski, Kevin Song, and Jeff Wooldridge for valuable comments and discussions. We thank SangMok Lee for excellent research assistance, and seminar participants at Michigan State, USC, U. Washington, the Canadian Econometrics Study Group meetings (2011, Toronto), and the Vanderbilt conference, Identification and Inference in Microeconometrics (2012) for useful comments.

<sup>2</sup>See, for example, Rosenbaum and Rubin (1983), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998), Hirano, Imbens, and Ridder (2003), Chernozhukov, Fernandez-Val, and Melly (2013), and Rothe (2012).

representations of the marginal and counterfactual distributions via an inverse propensity-score reweighting of the data and a continuous version of the classical monotone rearrangement inequality (see Hardy, Littlewood, and Polya (1934); Cambanis, Simons, and Stout (1976)), we obtain sharp bounds on the marginal and counterfactual distributions and policy parameters of interest, including average treatment effects (ATE) and average effects of treatment on the treated (ATT).

Recent work in the treatment effects literature has made use of the result in Cambanis, Simons, and Stout (1976) and inequalities bounding the distribution functions of a sum or difference between two random variables with fixed marginals in, for example, Frank, Nelsen, and Schweizer (1987) to evaluate distributional treatment effect parameters that depend on the joint distribution of the potential outcomes (such as the probability of a positive individual treatment effect and the median of the distribution of the individual treatment effect for the treated). These works include Fan and Park (2009, 2010, 2012), Firpo and Ridder (2009), Heckman, Smith, and Clements (1997), and Fan and Zhu (2009) who adopted the selection on observables assumption; and Fan and Wu (2010), which considered a class of latent threshold-crossing models. Unlike the current paper, however, these works assume that outcomes and covariates are observed in the same data set so that the *marginal* and counterfactual marginal distributions of potential outcomes are point identified. Extending these results, this paper establishes sharp bounds on the distributional treatment effect parameters that depend on the *joint* distribution of the potential outcomes when the marginals are partially identified.

The literature on data combination is much smaller. Manski (2000; esp. Section 5) considered bounds for a treatment effect model when the aggregate treatment outcomes and agent demographics are separately observed.<sup>3</sup> Cross and Manski (2002) derived sharp bounds on the “long regression” of a dependent variable  $Y$  on two sets of discrete covariates  $Z_1$  and  $Z_2$ , when only the conditional distributions of  $Y|Z_1$  and  $Z_2|Z_1$  are identified from separate datasets. Ridder and Moffitt (2007; Section 3.1) discussed the use of the Fréchet–Hoeffding inequality in data combination contexts.<sup>4</sup> Our main contribution here is to combine insights from the treatment effects literature with the monotone rearrangement inequality to obtain identification results for counterfactual distributions and treatment effects under data combination.

The rest of this paper is organized as follows. Section 2 introduces the modeling framework, some examples, and the unconfoundedness assumption. In Section 3, we present the main identification results. Section 4 concludes.

<sup>3</sup>The ecological inference literature also considers the partial identification problem when combining aggregate and individual-level data (e.g., Cross and Manski (2002) and references therein). The two-sample IV literature has considered instrumental variables models in which the outcome and the endogenous variables are observed in separate data sets (Klevmarken (1982)).

<sup>4</sup>For a reference on Fréchet–Hoeffding inequality, see Joe (1997).

Throughout the rest of this paper, we use  $F_{A|B}(\cdot|b)$  and  $f_{A|B}(\cdot|b)$  to denote the distribution function and density function of the random variable  $A$  conditional on  $B = b$ . For a distribution function  $F(\cdot)$ , we use  $F^{-1}(\cdot)$  to denote its quantile function.

## 2. THE MODELING FRAMEWORK AND ASSUMPTIONS

We now describe our treatment effects model, which follows closely the “potential outcomes” approach of Rubin (1974). We let  $D \in \{0, 1\}$  denote the two states of a binary treatment<sup>5</sup> and let  $Y_D$  denote the corresponding outcome variable of interest for  $D = 0, 1$ . Each individual agent is viewed as having both treatment and control outcomes  $Y_1$  and  $Y_0$ . Both  $Y_1$  and  $Y_0$  are potentially observed, but only one is actually observed. The actual observed outcome is  $Y \equiv Y_1D + Y_0(1 - D)$ . Let  $Z$  denote conditioning covariates (typically demographic variables) which can affect both  $D$  and  $(Y_0, Y_1)$ .

As a departure from the existing literature, we assume that the variables  $(Y, D, Z)$  are not observed in a single data set. Instead, we observe two separate data sets: (i) the *outcome* data set contains  $(Y, D)$ , while (ii) the *demographics* data set contains  $(Z, D)$ . We introduce several examples below.

**EXAMPLE A—Long-Run Returns to College Attendance:** This data problem arises naturally in situations when the outcome of interest is a long-run outcome which is not available immediately following the treatment. For example, consider the effect of college attendance on lifetime earnings, for which there is a very large existing empirical literature. Typically, long panels, like the PSID or NLSY, are used to assess the long-run returns to college. But recent papers using the National Longitudinal Survey of Adolescent Health (“Addhealth”) data set, which is a repeated cross-section of high school students, have uncovered many rich determinants of college attendance, including parental, classroom, and even genetic factors that are not measured in other data sets (see, e.g., Shanahan, Erickson, Vaisey, and Smolen (2008)).

Here  $Y$  denotes long-run earnings, observed in the PSID, while  $Z$  denotes specific determinants of college attendance, such as whether friends go to college, measures of parental attention, and genetic factors, which are only observed in Addhealth. The treatment variable  $D \in \{0, 1\}$  indicates whether a student attended college, and is observed in both the PSID and Addhealth.

**EXAMPLE B—Tax Payments Across Household Types:** For questions of tax incidence, data sets of individual tax returns are available. But tax returns contain very little demographic information on taxpayers. For instance, one may wish to examine how tax payments vary across household types—single households, couples without children, and households with children. Tax payments

<sup>5</sup>As in the examples below, these treatments can be policy interventions as well as different time periods.

and household type are observed from tax returns, but other demographic and labor market variables that are related to both tax payments and household type, such as years of education, occupational sector, and hours of work, are available in labor market data sets such as the *Current Population Survey*. In this example,  $Y$  denotes tax payments,  $D$  indexes the different household types, and  $Z$  are these additional demographic variables not observable from tax returns.

**EXAMPLE C—Changes in Wage Distribution Across Time:** This is an adaptation of DiNardo, Fortin, and Lemieux (1996, DFL). Here  $D$  is a binary indicator for two different years and  $Y_D$  denotes wages in year  $D$ . DFL focused on estimating  $f_{Y_0|D}(\cdot|1)$ , which they interpreted as the counterfactual density of wages “if individual attributes had remained at their year-1 levels but workers had been paid according to the year-0 wage schedule.”  $Z$  are additional covariates that affect wages. In DFL’s paper, all the variables are observed in the same data set, but in other applications, there are variables of interest (such as supply-side wage determinants, including industry profitability or import penetration<sup>6</sup>) which are not observed in the same data set as wages. In the latter case, our results here can be used to bound the counterfactual wage distributions.<sup>7</sup>

Next, we present the unconfoundedness or selection on observables assumption, which consists of two conditions. The first is a conditional independence assumption, while the second is an assumption about the support of the propensity score. Let  $\mathcal{Z}$  denote the support of  $Z$ .

- (C1) Let  $(Y_1, Y_0, D, Z)$  have a joint distribution. For all  $z \in \mathcal{Z}$ ,  $(Y_1, Y_0)$  is jointly independent of  $D$  conditional on  $Z = z$ .
- (C2) For all  $z \in \mathcal{Z}$ ,  $0 < p(z) < 1$ , where the propensity score  $p(z) = \Pr(D = 1|Z = z)$ .

**EXAMPLE A—Continued:** To understand the conditional independence assumption (C1) further, we consider a model  $Y_d = h(d, \eta_d)$  for  $d = 0, 1$  and  $D = I(\varepsilon > 0)$ , where  $\eta_d$  captures unobservables affecting earnings and  $\varepsilon$  captures unobservables affecting the schooling decision.  $Z$  are demographic variables affecting both  $\eta_d$  and  $\varepsilon$  such that  $\eta_d = Z + U_d$  and  $\varepsilon = Z + V$ , where  $Z$ ,  $U_d$ , and  $V$  are independent of each other.<sup>8</sup> Then the conditional indepen-

<sup>6</sup>For example, Feenstra and Hanson (2003).

<sup>7</sup>Fortin, Lemieux, and Firpo (2010) noted the formal equivalence between evaluating counterfactual distributions and evaluating treatment effects under the unconfoundedness assumption (Conditions (C1) and (C2)). See Chernozhukov, Fernandez-Val, and Melly (2013), and Rothe (2012) for related work.

<sup>8</sup>This is a simplified version of the setup considered in Carneiro, Hansen, and Heckman (2003) and Aakvik, Heckman, and Vytlačil (2005).

dence assumption holds. That is, college attendance is exogenous, conditional on  $Z$ , in the sense that, given the demographic variables in  $Z$ , the unobservables  $\eta_d$  do not affect college attendance. However, without conditioning on  $Z$ , the unobservables  $\eta_d$  do affect college attendance through  $Z$ . Without the demographics data set, existing papers in the literature have relied on instrumental variables to identify treatment effect parameters. This paper provides an alternative methodology when the demographics  $Z$  are available in a data set separately from the other variables.

*The Usual Approach*

When  $(Y, D, Z)$  are all observed in a single data set (so that there is no need for data combination), it is well known that, under (C1) and (C2), the marginal distributions  $F_{Y_1}(y)$ ,  $F_{Y_0}(y)$  and the counterfactual distribution function  $F_{Y_0|D}(y|1)$  are identified. Specifically,  $F_{Y_0|D}(y|1)$  is identified through

$$(2.1) \quad F_{Y_0|D}(y|1) = \int F_{Y_0|Z,D}(y|z, 1) dF_{Z|D}(z|1) \\ = \int F_{Y_0|Z,D}(y|z, 0) dF_{Z|D}(z|1),$$

in which the second equality holds under (C1);  $F_{Y_1}(y)$  and  $F_{Y_0}(y)$  are identified through

$$(2.2) \quad F_{Y_d}(y) = \int F_{Y_d|Z}(y|z) dF_Z(z) \\ = \int F_{Y_d|Z,D}(y|z, d) dF_Z(z) \quad \text{for } d = 0, 1.$$

Thus, parameters that are functionals of  $F_{Y_1|Z}(\cdot|z)$ ,  $F_{Y_0|Z}(\cdot|z)$ ,  $F_{Y_0|D}(\cdot|1)$ , including the ATE and ATT, are also identified.

However, when  $(Y, D)$  and  $(Z, D)$  are observed in separate data sets, we face a fundamental identification problem:  $F_{Y_d|Z,D}(y|z, d)$  is not point identified from the sample information, so it is easy to see from (2.1) and (2.2) that  $F_{Y_1|Z}(\cdot|z)$ ,  $F_{Y_0|Z}(\cdot|z)$ , and  $F_{Y_0|D}(\cdot|1)$  are not point identified. To tackle this problem, we make use of the alternative expressions for  $F_{Y_1}(y)$ ,  $F_{Y_0}(y)$ , and  $F_{Y_0|D}(y|1)$  in terms of inverse propensity-score weighted averages below:

$$(2.3) \quad F_{Y_1}(y) = E \left[ \frac{D}{p(Z)} I\{Y \leq y\} \right], \quad F_{Y_0}(y) = E \left[ \frac{1-D}{1-p(Z)} I\{Y \leq y\} \right],$$

$$(2.4) \quad F_{Y_0|D}(y|1) = \frac{1}{p_1} E \left[ \frac{(1-D)p(Z)}{1-p(Z)} I\{Y \leq y\} \right],$$

where  $p_1 = \Pr(D = 1)$ . The expectations in (2.3) and (2.4) are not point identified from the available data. We develop sharp bounds on these quantities in the next section.

### 3. IDENTIFYING TREATMENT EFFECTS UNDER DATA COMBINATION

In this section, we develop sharp bounds for the marginal and counterfactual marginal distributions of the potential outcomes  $Y_0, Y_1$  and for functionals of these distributions, including the traditional program evaluation parameters such as the ATE and ATT. We also demonstrate how sharp bounds on the marginal and counterfactual marginal distributions can be used to obtain sharp bounds on distributional treatment effects including the probability of a positive individual treatment effect and the median of the distribution of the individual treatment effect.

Our main identification results exploit a continuous version of the classical monotone rearrangement inequality in Hardy, Littlewood, and Polya (1934), a special case of Theorem 2 in Cambanis, Simons, and Stout (1976).<sup>9</sup> For convenience, we present it in the next lemma.

**LEMMA 3.1**—The Cambanis–Simons–Stout Inequality: *Let  $S$  and  $T$  denote two random variables with known marginal distribution functions  $F_S$  and  $F_T$ . Assume  $S$  and  $T$  have finite variances. Then*

$$\int_0^1 F_S^{-1}(1-u)F_T^{-1}(u) du \leq E(ST) \leq \int_0^1 F_S^{-1}(u)F_T^{-1}(u) du.$$

*The bounds are finite, and without additional information, they are sharp.*

It is worth pointing out that the Cambanis–Simons–Stout inequality provides sharp bounds on  $E(ST)$  when the marginal distributions of  $S, T$  are known, while an application of the Cauchy–Schwarz inequality to  $E(ST)$  in this case leads to bounds that are, in general, not sharp. Throughout the rest of this paper, we make the following assumption:

**ASSUMPTION (I):** *Let  $W = 1/p(Z)$  and  $V = 1/[1 - p(Z)]$ . Assume  $\text{Var}(W) < \infty$ ,  $\text{Var}(V) < \infty$ , and  $\text{Var}(V/W) < \infty$ . In addition, let  $g$  denote a measurable function such that  $\text{Var}(g(Y_d)) < \infty$  for  $d = 1, 0$ .*

This assumption, along with Lemma 3.1 above, assures the finiteness of the bounds we will consider below. A simple sufficient (but by no means necessary) condition for the finiteness of the variances of  $W, V$ , and  $V/W$  in the previous

<sup>9</sup>See also Chernozhukov, Fernandez-Val, and Galichon (2010) for a recent application of monotone rearrangement to constructing quantile curves without crossing.

assumption is that the propensity score  $p(z)$  is bounded away from zero and unity on the support of  $Z$ . For example, if a probit model is used to model the propensity score, then these variances are finite if  $Z$  has bounded support.

### 3.1. A General Result

Our first series of results establishes sharp bounds on the mean of  $g(Y_d)$ .

**THEOREM 3.2:** (i) *Let  $\mu_d(g) \equiv E(g(Y_d))$ . Then  $\mu_d^L(g) \leq \mu_d(g) \leq \mu_d^U(g)$ , for  $d = 1, 0$  and*

$$\begin{aligned} \mu_1^L(g) &= E \left[ D \int_0^1 F_{g(Y)|D}^{-1}(1-u|D) F_{W|D}^{-1}(u|D) du \right], \\ \mu_1^U(g) &= E \left[ D \int_0^1 F_{g(Y)|D}^{-1}(u|D) F_{W|D}^{-1}(u|D) du \right], \\ \mu_0^L(g) &= E \left[ (1-D) \int_0^1 F_{g(Y)|D}^{-1}(1-u|D) F_{V|D}^{-1}(u|D) du \right], \\ \mu_0^U(g) &= E \left[ (1-D) \int_0^1 F_{g(Y)|D}^{-1}(u|D) F_{V|D}^{-1}(u|D) du \right]. \end{aligned}$$

*Without additional information, the bounds are sharp.*

(ii) *Let  $\mu_{d|1}(g) \equiv E(g(Y_d)|D = 1)$ . Then  $\mu_{1|1}(g)$  is identified:  $\mu_{1|1}(g) = E(Dg(Y))/p_1$  and  $\mu_{0|1}^L(g) \leq \mu_{0|1}(g) \leq \mu_{0|1}^U(g)$ , where*

$$\begin{aligned} \mu_{0|1}^L(g) &= \frac{1}{p_1} E \left[ (1-D) \int_0^1 F_{g(Y)|D}^{-1}(1-u|D) F_{V/W|D}^{-1}(u|D) du \right], \\ \mu_{0|1}^U(g) &= \frac{1}{p_1} E \left[ (1-D) \int_0^1 F_{g(Y)|D}^{-1}(u|D) F_{V/W|D}^{-1}(u|D) du \right]. \end{aligned}$$

*Without additional information, the bounds are sharp.*

**PROOF:** Consider  $\mu_1(g)$ . An analogue of Eq. (2.3) gives us an expression for  $\mu_1(g)$  in terms of the variables  $(Y, D, Z)$ , but we cannot compute this because we do not observe the joint distribution  $(Y, D, Z)$ , but only the two separate distributions of  $(Y, D)$  and  $(D, Z)$ . The data set on  $(D, Z)$  allows us to identify the propensity score  $p(z)$ . Then, rearranging the expression, we get

$$\mu_1(g) = E \left[ \frac{D}{p(Z)} g(Y) \right] = E(Dg(Y)W) = E(DE[g(Y)W|D]).$$

The rightmost quantity here contains the term  $E[g(Y)W|D]$ , which is the (conditional) expectation of a product of two random variables  $g(Y)$  and  $W$ , which are observed in different data sets, so that the expectation cannot be computed

feasibly. However, we can apply Lemma 3.1 to obtain bounds on the expectation of their product. This leads to the bounds for  $\mu_1(g)$  in part (i) of Theorem 3.2. Similarly, by using the expression  $\mu_0(g) = E[\frac{1-D}{1-p(Z)}g(Y)]$ , we obtain the bounds for  $\mu_0(g)$  in part (i). For part (ii), noting that  $V/W = p(Z)/[1 - p(Z)]$ , we get  $p_1\mu_{0|1}(g) = E[(1 - D)(\frac{V}{W})g(Y)]$  and the bounds in part (ii).

The bounds for  $\mu_1(g)$  are sharp, in that there exist distributions of  $(D, Y, W)$  which attain these bounds. The upper bound on  $\mu_1(g)$  is achieved when, conditional on  $D$ ,  $(g(Y), W)$  are perfectly positively dependent on each other; the lower bound is achieved when, conditional on  $D$ ,  $(g(Y), W)$  are perfectly negatively dependent on each other. Analogously, the upper bound on  $\mu_0(g)$  is achieved when, conditional on  $D$ ,  $(g(Y), V)$  are perfectly positively dependent on each other and the lower bound is achieved when, conditional on  $D$ ,  $(g(Y), V)$  are perfectly negatively dependent on each other. Q.E.D.

$\mu_1^L(g)$  and  $\mu_1^U(g)$  are identified from the sample information, as  $F_{g(Y)|D}(\cdot|d)$  is identified from the first data set,  $F_{W|D}(\cdot|d)$  ( $F_{V|D}(\cdot|d)$ ) is identified from the second data set, and the expectation in the expressions for  $\mu_1^L(g)$  and  $\mu_1^U(g)$  can be identified from either data set (or both).

### 3.2. Counterfactual Distributions and Treatment Effects

Let  $\Delta \equiv Y_1 - Y_0$  denote the individual treatment effect. Let  $\mu_\Delta$  and  $\mu_{\Delta|1}$  denote, respectively, the ATE and the ATT, that is,  $\mu_\Delta = E(\Delta)$  and  $\mu_{\Delta|1} = E(\Delta|D = 1)$ . Bounds on  $\mu_\Delta$  and  $\mu_{\Delta|1}$  follow immediately from Theorem 3.2:

$$(3.1) \quad \mu_1^L - \mu_0^U \leq \mu_\Delta \leq \mu_1^U - \mu_0^L \quad \text{and} \\ \frac{1}{p_1}E[DY] - \mu_{0|1}^U \leq \mu_{\Delta|1} \leq \frac{1}{p_1}E[DY] - \mu_{0|1}^L.$$

Let  $g(Y_d) = I\{Y_d \leq y\}$  in Theorem 3.2. Noting that

$$F_{I_Y|D}^{-1}(u|D) = \begin{cases} 0, & \text{for } u \in [0, 1 - F_{Y|D}(y|D)), \\ 1, & \text{for } u \in [1 - F_{Y|D}(y|D), 1], \end{cases}$$

where  $I_Y = I\{Y \leq y\}$ , we obtain bounds for  $F_{Y_1}(y)$ ,  $F_{Y_0}(y)$  in part (i) of Theorem 3.3 below. Bounds for the counterfactual marginal distribution function  $F_{Y_0|D}(y|1)$  are obtained similarly.

**THEOREM 3.3:** (i) For  $d = 0, 1$ , we have  $F_d^L(y) \leq F_{Y_d}(y) \leq F_d^U(y)$ , where

$$F_1^L(y) = E \left[ D \int_0^{F_{Y_1|D}(y|D)} F_{W|D}^{-1}(u|D) du \right], \\ F_1^U(y) = E \left[ D \int_{1-F_{Y_1|D}(y|D)}^1 F_{W|D}^{-1}(u|D) du \right],$$

$$F_0^L(y) = E \left[ (1 - D) \int_0^{F_{Y|D}(y|D)} F_{V|D}^{-1}(u|D) \, du \right],$$

$$F_0^U(y) = E \left[ (1 - D) \int_{1-F_{Y|D}(y|D)}^1 F_{V|D}^{-1}(u|D) \, du \right].$$

Without additional information, the bounds are sharp (both pointwise and uniformly).

(ii)  $F_{Y_1|D}(y|1)$  is identified:  $F_{Y_1|D}(y|1) = E[DI\{Y \leq y\}]/p_1$ , and  $F_{Y_0|D}(y|1)$  is partially identified:  $F_{0|D}^L(y|1) \leq F_{Y_0|D}(y|1) \leq F_{0|D}^U(y|1)$ , where

$$F_{0|D}^L(y|1) = \frac{1}{p_1} E \left[ (1 - D) \int_0^{F_{Y_1|D}(y|D)} F_{V/W|D}^{-1}(u|D) \, du \right] \quad \text{and}$$

$$F_{0|D}^U(y|1) = \frac{1}{p_1} E \left[ (1 - D) \int_{1-F_{Y_1|D}(y|D)}^1 F_{V/W|D}^{-1}(u|D) \, du \right].$$

Without additional information, the bounds are sharp (both pointwise and uniformly).

We note that the distribution bounds in Theorem 3.3 are not only pointwise sharp but also uniformly sharp, that is, the upper and lower bounds are distribution functions that are attainable for specific data-generating processes. To see this, consider the bounds on  $F_1(\cdot)$ . Both  $F_1^L(\cdot)$  and  $F_1^U(\cdot)$  are distribution functions.  $F_1^L(\cdot)$  is the distribution function of  $Y_1$  when, conditional on  $D$ ,  $I_Y$  and  $W$  are perfectly negatively dependent on each other, or, equivalently,  $Y$  and  $W$  are perfectly positively dependent on each other; the upper bound  $F_1^U(\cdot)$  is the distribution function of  $Y_1$  when, conditional on  $D$ ,  $Y$  and  $W$  are perfectly negatively dependent on each other.

The uniform sharpness of the bounds in Theorem 3.3 allows us to establish sharp bounds on monotone functionals of the marginal or counterfactual marginal distribution functions. Such functionals include the quantile treatment effects (QTE) defined as

$$QTE_\tau = F_{Y_1}^{-1}(\tau) - F_{Y_0}^{-1}(\tau) \quad \text{and}$$

$$QTE_{\tau|1} = F_{Y_1|D}^{-1}(\tau|1) - F_{Y_0|D}^{-1}(\tau|1), \quad \tau \in (0, 1).$$

For instance, we have that  $F_d^{U^{-1}}(\tau) \leq F_{Y_d}^{-1}(\tau) \leq F_d^{L^{-1}}(\tau)$ , yielding the bounds on  $QTE_\tau$  as  $F_1^{U^{-1}}(\tau) - F_0^{L^{-1}}(\tau) \leq QTE_\tau \leq F_1^{L^{-1}}(\tau) - F_0^{U^{-1}}(\tau)$ .

To see explicitly how these sharp bounds can be achieved, consider the stated upper bound for  $QTE_\tau$ , namely,  $F_1^{L^{-1}}(\tau) - F_0^{U^{-1}}(\tau)$ . It follows from the discussion after Theorem 3.3 that this upper bound is achieved when, conditional on  $D$ ,  $Y$  and  $W$  are perfectly positively dependent on each other and  $Y$  and  $V$  are

perfectly negatively dependent on each other. By the definitions of  $W$  and  $V$ , we conclude that this upper bound is achieved when, conditional on  $D$ ,  $Y$  and  $P(Z)$  are perfectly negatively dependent on each other.

### 3.3. Distributional Treatment Effects

Under the selection on observables assumption, when the outcomes and covariates are observed in the same data set, Fan and Park (2009, 2010) have established bounds on the distribution of the individual treatment effect and the distribution for the treated:

$$F_{\Delta}(\delta) = \Pr(\Delta \leq \delta) \quad \text{and} \quad F_{\Delta}(\delta|D=1) = \Pr(\Delta \leq \delta|D=1).$$

These are useful when one is interested in distributional treatment effects such as the probability of a positive individual treatment effect: either  $\Pr(\Delta > 0)$  or  $\Pr(\Delta > 0|D=1)$ , and the median of  $\Delta$ . Theorem 3.3 and the lemma below, adapted from Frank, Nelsen, and Schweizer (1987), allow us to establish similar results to Fan and Park (2009, 2010, 2012) in our context.

LEMMA 3.4: *Let  $S$  and  $T$  denote two random variables with fixed marginal distribution functions  $F_S$  and  $F_T$ . Further let  $F_{S-T}(\delta)$  denote the distribution function of  $(S - T)$ . Then  $F_{S-T}^L(\delta) \leq F_{S-T}(\delta) \leq F_{S-T}^U(\delta)$ , where*

$$F_{S-T}^L(\delta) = \max\left(\sup_y [F_S(y) - F_T(y - \delta)], 0\right),$$

$$F_{S-T}^U(\delta) = 1 + \min\left(\inf_y [F_S(y) - F_T(y - \delta)], 0\right).$$

Consider, for instance, the distribution function  $F_{\Delta}(\delta|D=1)$ . From Theorem 3.3 and the conditional version of Lemma 3.4, we have

$$F_{\Delta}^L(\delta|D=1) \leq F_{\Delta}(\delta|D=1) \leq F_{\Delta}^U(\delta|D=1), \quad \text{where}$$

$$F_{\Delta}^L(\delta|D=1) = \max\left(\sup_y [F_{Y_1|D}(y|1) - F_{0|D}^U(y - \delta|1)], 0\right),$$

$$F_{\Delta}^U(\delta|D=1) = 1 + \min\left(\inf_y [F_{Y_1|D}(y|1) - F_{0|D}^L(y - \delta|1)], 0\right).$$

Sharp bounds on the quantile function of  $F_{\Delta}(\delta|D=1)$  follow directly from sharp bounds on  $F_{\Delta}(\delta|D=1)$ .

## 4. CONCLUDING REMARKS

We consider the identification of counterfactual distributions and treatment effects when the outcome variables and conditioning covariates are observed

in separate data sets. Even under the selection on observables assumption, the marginal and counterfactual marginal distributions (hence the average treatment effect parameters) are no longer point identified, and we utilize the monotone rearrangement inequality to derive sharp bounds on the counterfactual distribution and policy parameters of interest. While this note focuses exclusively on identification, a companion paper (Fan, Sherman, and Shum (2012)) considers inference in these models and includes an empirical application to predicting counterfactual voting outcomes in U.S. elections.

Extensions of the results in this note to the case that the separate data sets contain a common covariate  $X$ , that is, one data set contains observations on  $(Y, D, X)$  and the other contains  $(D, Z, X)$ , are straightforward.

#### REFERENCES

- AAKVIK, A., J. HECKMAN, AND E. VYTLACIL (2005): "Treatment Effects for Discrete Outcomes When Responses to Treatment Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs," *Journal of Econometrics*, 125, 15–51. [814]
- CAMBANIS, S., G. SIMONS, AND W. STOUT (1976): "Inequalities for  $Ek(X, Y)$  When the Marginals Are Fixed," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 36, 285–294. [812,816]
- CARNEIRO, P., K. T. HANSEN, AND J. HECKMAN (2003): "Estimating Distributions of Treatment Effects With an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44 (2), 361–422. [814]
- CHERNOZHUKOV, V., I. FERNANDEZ-VAL, AND A. GALICHON (2010): "Quantile and Probability Curves Without Crossing," *Econometrica*, 78, 1093–1125. [816]
- CHERNOZHUKOV, V., I. FERNANDEZ-VAL, AND B. MELLY (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81 (6), 2205–2268. [811,814]
- CROSS, P. J., AND C. F. MANSKI (2002): "Regressions, Short and Long," *Econometrica*, 70 (1), 357–368. [812]
- DINARDO, J., N. FORTIN, AND T. LEMIEUX (1996): "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044. [814]
- FAN, Y., AND S. PARK (2009): "Partial Identification of the Distribution of Treatment Effects and Confidence Sets," in *Advances in Econometrics: Nonparametric Econometric Methods*. Bingley, U.K.: Emerald Group. [812,820]
- (2010): "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory*, 26, 931–951. [812,820]
- (2012): "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments," *Journal of Econometrics*, 167, 330–344. [812,820]
- FAN, Y., AND J. WU (2010): "Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and Its Confidence Sets," *Review of Economic Studies*, 77, 1002–1041. [812]
- FAN, Y., AND D. ZHU (2009): "Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes," Working Paper, Vanderbilt. [812]
- FAN, Y., R. SHERMAN, AND M. SHUM (2012): "Partial Identification of Treatment Effects Under Data Combination: Inference and an Application to Counterfactual Election Prediction," Working Paper. [821]
- FEENSTRA, R., AND G. HANSON (2003): "Global Production Sharing and Rising Inequality: A Survey of Trade and Wages," in *Handbook of International Trade*. Oxford: Blackwell Publishing. [814]
- FIRPO, S., AND G. RIDDER (2009): "Bounds on Functionals of the Distribution of Treatment Effects," Working Paper. [812]

- FORTIN, N., T. LEMIEUX, AND S. FIRPO (2010): "Decomposition Methods in Economics," in *Handbook of Labor Economics*, Vol. 4, ed. by D. Card and O. Ashenfelter. New York: North-Holland. [814]
- FRANK, M., R. NELSEN, AND B. SCHWEIZER (1987): "Best-Possible Bounds on the Distribution of a Sum—A Problem of Kolmogorov," *Probability Theory and Related Fields*, 74, 199–211. [812,820]
- HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [811]
- HARDY, G., J. LITTLEWOOD, AND G. POLYA (1934): *Inequalities*. Cambridge: Cambridge University Press. [812,816]
- HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098. [811]
- HECKMAN, J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535. [812]
- HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [811]
- JOE, H. (1997): *Multivariate Models and Multivariate Dependence Concepts*. London: Chapman & Hall. [812]
- KLEVMARKEN, N. A. (1982): "Missing Variables and Two-Stage Least-squares Estimation From More Than One Dataset," in *1981 Proceedings of the American Statistical Association*, Business and Economic Statistics Section, 156–161. [812]
- MANSKI, C. F. (2000): "Identification Problems and Decision Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415–442. [812]
- RIDDER, G., AND R. MOFFITT (2007): "Econometrics of Data Combination," in *The Handbook of Econometrics*, Vol. 6B, Chapter 75. New York: North-Holland. [812]
- 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. [811]
- ROTHE, C. (2012): "Partial Distributional Policy Effects," *Econometrica*, 80, 2269–2301. [811,814]
- RUBIN, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [813]
- SHANAHAN, M., L. ERICKSON, S. VAISEY, AND A. SMOLEN (2008): "Environmental Contingencies and Genetic Propensities: Social Capital, Educational Continuation, and Dopamine Receptor Gene DRD2," *American Journal of Sociology*, 114 (Suppl.), S260–S286. [813]

*Dept. of Economics, University of Washington, Box 353330, Seattle, WA 98195, U.S.A.; fany88@u.washington.edu,*

*HSS, Caltech, 1200 East California Blvd., Pasadena, CA 91125, U.S.A.; sherman@hss.caltech.edu,*

*and*

*HSS, Caltech, 1200 East California Blvd., Pasadena, CA 91125, U.S.A.; mshum@hss.caltech.edu.*

*Manuscript received February, 2012; final revision received October, 2013.*